

Don't underestimate the enemy

When an infectious disease appears to be in decline, the agent that causes it tends to disappear from the biomedical research agenda. As recent events have revealed, that can be a mistake.

The outbreak of poliomyelitis that hit the Caribbean island of Hispaniola last summer took everyone by surprise. The news that weakened, live viruses in the oral polio vaccine had reverted to a form that can initiate a disease outbreak is forcing the World Health Organization to re-examine the final stages of its global campaign to eradicate polio (see page 278).

Thankfully, the outbreak amounted to no more than eight cases. But it has exposed the limits of our knowledge about the polio virus. The genetic determinants of its transmissibility, for instance, remain obscure. And they seem likely to remain so, with polio now a low priority for the world's biomedical research agencies, which view it as a 'conquered' disease.

This attitude is short-sighted, because viruses have a habit of producing unpleasant surprises. In 1996 and 1997, for instance, as civil war raged in what is now the Democratic Republic of the Congo, the country was hit by a substantial outbreak of monkeypox. In Central Africa, contact with primates means that people do catch this disease from time to time. But the severity of the outbreak sparked fears that the virus had mutated into a more pathogenic form, or had acquired an enhanced ability to spread from person to person.

In the event, this appeared not to have happened. But some virologists argue that the threat is real, as the global eradication of smallpox — completed more than two decades ago — has left the door open for the related monkeypox virus to evolve and fill its vacant niche. Once more, however, testing this theory does not

seem to be high on the list of biomedical research priorities.

The last official samples of smallpox reside in secure facilities in the United States and Russia. But other stocks may exist in undeclared bioweapons laboratories — and the threat of biowarfare or bioterrorism provides another reason to maintain a broad view of the potential threat posed by viral diseases.

This was underlined last week by the revelation that a team in Australia had inadvertently created an unusually lethal form of mousepox. The researchers were trying to create a contraceptive vaccine, using a benign mousepox virus to express genes for proteins carried on mouse eggs. They also added the gene for interleukin-4 (IL-4), in an attempt to boost the antibody response against these proteins. But a paper in the latest issue of the *Journal of Virology* (75, 1205–1210; 2001) describes how the IL-4 gene also shut down the cellular arm of the animals' immune systems. As a result, the virus killed the vaccinated mice within days. Such findings show that it may not take much for a malevolent bioengineer to create a hideously effective weapon.

Advances in fields from immunology to epidemiological modelling mean that we are in a better position than ever to respond to the threats posed by viral evolution and bioterrorism. High-throughput gene sequencing, in particular, means that a virus's genetic secrets can be rapidly laid bare. But deploying these tools effectively may require a strategy of expecting the unexpected. Apparently conquered or benign viruses cannot simply be ignored. ■

The stuff of ideas

Introducing new ways in which researchers and others can stimulate *Nature's* readers.

This issue sees the launch of two series of essays, devoted respectively to writing and concepts in science. It also contains an unusual collection of overviews in the Insight section which focuses not on the current state of research but, in a more visionary spirit, on its future (see page 385). All of this represents an effort to focus on aspects of the essence of science, rather than on new results or new developments in science policy. And at some level, all focus on something the literature sees too little of: scientific ideas.

The self-censorship of ideas from scientific journals is, up to a point, desirable. The value of ideas as perceived by others (including referees) can be highly subjective. Few have the time to invest in the hope that an idea has potential, only to find that it does not. And an idea's true originality can be difficult to evaluate. New ideas, in short, are usually best cultivated in private, especially if one wants to reap the rewards that their fulfilment might bring. And the literature is burgeoning enough already.

Nevertheless, open discussion of ideas seems undervalued. Occasionally, an old idea may deserve public re-examination because it has more to offer or is being misrepresented — as a result, perhaps, of the 'Chinese whispers' of second- or *n*th-hand citation. Or a recent idea may deserve explication to a wide audience to fulfil its potential between disciplines. Alternatively, an idea

may deserve examination for its development.

It is with these considerations in mind that the series Concepts (see page 289) has been conceived. Some scientists nigger away at a concept for years, finding it an ever-fruitful source of stimulation or puzzlement. Some may be exasperated by others' perceived misuse of a concept. Either way, they have a point of view that deserves to be captured for the rest of us. Each article will therefore not only communicate some aspect of a concept to readers, but will also bring a touch of the author's personal perspective along with it.

The act of writing itself, in and about science, also deserves celebration and analysis. Hence our other new section, Words (see page 287). The weekly articles will explore issues, anecdotes and episodes (both historical and contemporary) that elucidate the relationship between science and words. We shall be publishing contributions from authors in the many disciplines that touch on science, words and language, including leading linguists, historians of science, poets, playwrights, novelists, anthropologists and scientists who are themselves excellent writers.

These articles are essays — a word defined by at least one dictionary as "a literary composition". We hope that, in both new series, the writing itself will be found to be as enjoyable as the messages are refreshing. ■





Root fixation
Geneticists set to
unravel link between
legumes and bacteria
p272



Animal alarm
Questions raised
over antibiotic use
in livestock
p273



Listen up
Scientists gain the
ear of Spain's prime
minister
p274



Meat misery
Germany faces
political crisis over
BSE fears
p275

Harvard to spend 'hundreds of millions' on further expansion

Rex Dalton

Harvard University is planning to invest hundreds of millions of dollars in a new natural history museum, an expanded faculty and extra facilities for scientific research and teaching.

The cost has not been determined, but some estimates put the total at over \$700 million. Jeremy Knowles, dean of the faculty of arts and sciences, says the cost "would be in the hundreds of millions of dollars over the next five years". The move follows a \$200 million investment in new science institutes announced just two years ago.

Investment profits from Harvard's \$19.2 billion endowment — which earned a 32% return for the year ending last June — will be used to finance the undertaking, officials say.

Rather than "just running on a richer mixture", says Knowles, Harvard plans to "grasp the opportunity to improve the quality of the undergraduate experience and the lives of graduate students and faculty".

During the next decade, six new chairs will be added to the faculty each year. The 60 new positions will be spread across all disciplines in the Faculty of Arts and Sciences, which currently has 650 professors, 3,500 graduate students and 6,500 undergraduates.

Several extra buildings for teaching and research will be built in the North Yard section of Harvard's campus at Cambridge, Massachusetts, next to the existing Museum of Comparative Zoology.

The new natural history museum, whose name and site have yet to be chosen, will put collections from several of Harvard's existing museums on public display, and the university's various scholarly collections will be updated and improved, Harvard officials say. The university's popular Glass Flowers collection of more than 4,000 botanical models will be a central feature in the new museum. Collections will also be housed there from the zoology museum, the Peabody Museum of Archaeology and Ethnology, the Mineralogical and Geological Museum, and the Botanical Museum.

The new museum and the updating of Harvard's scholarly collections are seen as an important contribution to the overall health



Solid backing: Harvard will fund growth using profits from its \$19.2 billion endowment fund.

of single-organism biology in the United States. "This all underscores the exciting time, showing the renewed potency of natural history museums and their collections," says James Hanken of Harvard's department of organismic and evolutionary biology.

The new buildings and areas vacated by museum collections, will house new laboratories, teaching areas and offices.

The plan comes two years after Harvard announced a major investment in facilities to house new institutes in genomics, nanotechnology, computer science and climate research. The Bauer Center for Genomic Research is under construction, and plans are being drawn up for the Center for Imaging and Mesoscale Structures.

► <http://www.harvard.edu>

Biotech sector still looking lively

Sally Lehrman, San Francisco

The Nasdaq high-technology stock market has just finished its worst year ever, and biotechnology stocks have been hit particularly hard over the past few weeks. But raising money for biotechnology ideas has never been easier, visitors to the sector's annual financial jamboree were told.

According to Wall Street analysts at the J. P. Morgan H&Q Healthcare conference in San Francisco last week, opportunities to develop new ideas in biotech are as strong as ever. Venture funding is robust, and commercial companies with cash available are also looking to invest.

Even though biotechnology stocks fell towards the end of 2000, many companies had used the positive market of the previous 12 months to strengthen their financial

position. Biotechnology concerns raised an estimated \$35 billion in total during 2000, making it a record year.

In the third quarter of 2000, venture capitalists put \$708 million into biotechnology companies, according to Pricewaterhouse Coopers: 150% more than in the same period of 1999.

The meeting also heard that a new subsector is emerging: 'post-genomics' companies, developing analytical techniques and high-speed technology to accelerate the translation of genomics into drug therapies. Robert Olan, a biotechnology analyst for J. P. Morgan H&Q in New York, said that universities have become more sophisticated about licensing agreements, and competition for their inventions has increased.



Rooting for success: *Lotus japonicus* benefits from the bacteria living in its root nodules (inset).

Japanese legume project may help to fix nitrogen problem

David Cyranoski, Tokyo

The mysterious relationship between legume-type plants and the nitrogen-fixing bacteria they harbour is to be investigated in a Japanese genome-sequencing project.

The nitrogen-grabbing function of the bacteria, which act in beans and other legumes, is of particular interest to researchers because it might replace high-nitrogen fertilizers as a means of accelerating plant growth. And nitrogen pollution of land and waterways is emerging as a major problem in Europe, east Asia and other intensively farmed regions of the world.

Japan's Kazusa DNA Research Institute in Chiba, a small agricultural district outside Tokyo, has already sequenced the genome of one bacterium involved, *Mesorhizobium loti*. It is now taking on the task of sequencing the associated legume, *Lotus japonicus*.

Satoshi Tabata, who heads the project, believes that unravelling the two genomes will lead to a better understanding of nitrogen fixation — the conversion of inorganic nitrogen from soil into organic nitrogen that plants can use for growth. Researchers may eventually use the genetic information to optimize nitrogen fixation in legumes, and perhaps apply it to other plants.

Plants cannot themselves fix nitrogen, but legumes and some other plant varieties have developed a symbiotic relationship with bacteria that can absorb nitrogen from the soil. The bacteria receive shelter and nutrients from the plant in exchange for nitrogen.

Researchers believe the fixation process involves several steps. First, the plant and bacterium undergo a recognition process, in which each notes the other's presence by

emitting chemical signals. The plant reacts to the bacterial signal by developing root 'nodules'. Once the bacteria have been taken up into the specialized nodule cells, they begin nitrogen fixation.

Tabata says the project will help to identify the genes that enable the plant and bacterium to recognize each other and the plant genes that control nodule formation.

Ultimately, some of these genes could be used in unrelated plants, such as corn and wheat, spurring 'natural' nitrogen fixation and alleviating the need for the current heavy use of high-nitrogen fertilizers.

Tabata's approach to the sequencing problem is to focus on sequences in the genome that are similar to genes of known function. This will provide useful information even before the genome is complete, he says. Jens Stougaard of the Laboratory of Gene Expression at the University of Aarhus in Denmark, which studies gene function by examining mutated versions of *L. japonicus* and other plants, says Kazusa's work is "laying out the path for the rest of us to follow".

Furthermore, according to Tabata, *L. japonicus* could serve as a model for other legumes that are agriculturally important but difficult to study. Its genome is only about 450 megabases, about two-fifths the size of its more celebrated cousin, the soy bean. And its fast maturation and small size make it a convenient experimental plant.

But at the current rate it could take 15 years to complete the sequence. The project is funded by the Chiba regional government. Tabata says he would like other groups to get involved, but finds most national governments want to focus directly on crop plants. ■

South Africa reveals funding allocations for research

Michael Cherry, Cape Town

South Africa has announced how it will split university research funding for nine focus areas, which it selected last year. The areas were chosen in an effort to tie research more closely to societal needs.

The allocations (see table) are expected to form the basis for funding by the National Research Foundation (NRF) over the next five years, although grants will be awarded on an annual basis.

Half of this year's NRF budget of R162 million (\$20.5 million) will go to peer-reviewed grants for established researchers. The rest will be allocated to other programmes such as the development of new research capacity at technical colleges and historically black universities.

Gerhard von Gruenewaldt, vice-president of the NRF, says that funds were allocated to reflect the value of grant applications received in each area. This 'bottom-up' approach ensures that the government is not dictating a research agenda, he says.

Not all researchers in the country are thrilled with the outcome, and von Gruenewaldt admits that the available funding will support only 70% of the projects that the agency wanted to approve. "This presents us with the challenge of raising a substantial amount of additional resources over the next few years to meet these demands," he says.

But some researchers complain that the NRF reallocated their proposals into the wrong focus area. Ed Rybicki, a microbiologist at the University of Cape Town, says that this happened to him, and that his proposal was consequently underfunded. "I think the whole process is arbitrary and badly thought out," he says. ■

	NRF FOCUS AREAS 2001–2005	Rand million
1	Strategic knowledge	16.5
2	Distinct South African research opportunities	5
3	Economic competitiveness	21
4	Poverty eradication	8
5	Ecological conservation	12.5
6	Information technology	4
7	Socio-political impact of globalization	1
8	Education	5
9	Indigenous knowledge	10

Hard focus: the areas were chosen to tie research to South Africa's social needs.

Group urges survey of antibiotics in animals

Meredith Wadman, Washington

The Union of Concerned Scientists (UCS) is calling for widescale government monitoring of the use of antibiotics in livestock. In a report released last week, it says the amounts used far exceed industry estimates and that widespread use is putting human health at risk by encouraging the development of resistant bacteria.

The report estimates that 70% of all antibiotics used in the United States are fed to healthy animals. It urges the Food and Drug Administration (FDA) and the US Department of Agriculture to compel the drug and livestock industries to report the quantities, dosages and treatment periods of all antimicrobials used on food animals.

"Which antibiotics are most in use? Have use patterns changed over time? Why? These are all very basic questions, and in the US, you can't answer them. There are no data compiled by the government either on production or use," says Margaret Mellon, the report's lead author.

The UCS, which represents about 50,000 US scientists and citizens, timed its report to influence an FDA meeting on 22 January on antimicrobial use in livestock. The report says that, without thorough and reliable data, it is impossible to combat growing antimicrobial resistance in human pathogens.

Industry has in the past resisted providing such data. Livestock and poultry producers have used low-level antibiotics non-thera-

peutically in animals for decades to promote growth and prevent disease. They argue that restrictions would put meat prices up and compromise food safety.

The UCS report uses publicly available data to arrive at an admittedly approximate estimate that 24.6 million pounds (11,200 tonnes) of antimicrobials are fed to US poultry, pigs and cattle annually for non-therapeutic uses, an increase of 50% since 1985.

Industry sources dispute that number. The Animal Health Institute (AHI) in Washington, which represents the makers of animal drugs, estimated last year that 8,100 tonnes of antibiotics were used for all purposes in animals, of which only 4,200

tonnes are antibiotics also used in humans.

"The assumption is that somehow human health is being compromised," says John Keeling, who spearheaded the AHI survey. "That has not been demonstrated."

But Stephen Sundlof, director of the FDA's Center for Veterinary Medicine, says that transmission of antimicrobial-resistant pathogens from animals to humans in the food supply is an established fact, although the degree to which it occurs and the severity of the disease it causes remains in dispute.

Sundlof says that the agency has drafted regulations that will compel companies to make available most of the information that the UCS is calling for. ■



Pigging out? The widespread use of antibiotics on healthy animals might be affecting human health.

French museum report sparks researchers' revolt

Declan Butler, Paris

Researchers at the National Museum of Natural History in Paris are up in arms over a report on the future of research at the museum. The report, which is likely to determine the museum's research strategy, implies that biophysics and Earth sciences have no place in the institution.

Research should be centred around nine themes, the report says. These range from palaeontology to developmental biology, but

all involve studying biodiversity from an evolutionary perspective. The report was commissioned from an outside panel by the museum's interim administrator, Jean-Claude Moreno, as part of a plan to reform the museum (see *Nature* 401, 104; 1999).

Museum staff say they were initially optimistic, but that the process has proceeded without their consultation. Trade-union representatives and professors walked out of a council meeting at the museum on 15 December after reading out a statement to Moreno saying the report inspired "deception rather than hope".

One researcher says the reforms are being managed in an atmosphere of "extreme secrecy" and that there is "enormous apathy and exhaustion" among museum staff.

Moreno has promised that researchers will have the opportunity later in January to meet the committee that wrote the report, and hopes that most elements of the reform plan will be put in place before the summer.

Thérèse Garestier, director of the

biophysics lab, which the report acknowledges as a centre of excellence, says she does not understand how its work in functional genomics can be dissociated from biodiversity. She says that implementing the report would reduce her laboratory's role to servicing the museum's taxonomists.

Over 100 researchers from the four labs making up the Earth sciences department signed a motion deploring the assertion that their research has no place in the museum. The study, they say, "demonstrates a curious interpretation of the natural sciences... with no reference to [the] past, nor to the Earth on which life developed and evolved".

The report gives top priority to securing the housing and computerization of the museum's millions of specimens.

Paul Henderson, director of science at London's Natural History Museum and the only foreign member of the committee, says the report is a great opportunity for the museum to "re-establish itself as one of the great world centres in biodiversity". ■



Looking ahead: the National Museum of Natural History in Paris is focusing on biodiversity.

Genes may solve hormone-disrupter debate

Robert Triendl, Tokyo

Disagreements over the health effects of so-called endocrine-disrupting chemicals could eventually be resolved by studying the genetic variability of human and animal reactions to chemicals, researchers believe.

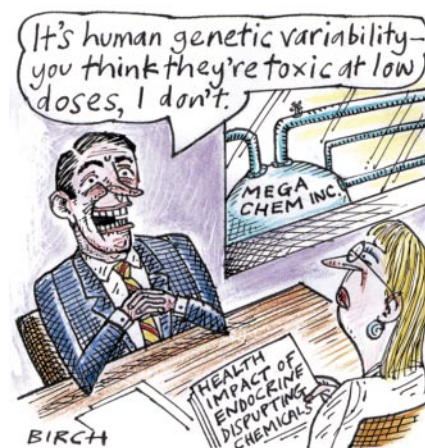
An international symposium on endocrine disrupters in Yokohama last month revealed a growing consensus that some of these chemicals can produce toxic effects at surprisingly low doses. But the meeting, the third in an annual series sponsored by the Japanese government's Environment Agency, found that the mechanisms underlying such effects are still poorly understood.

Endocrine disrupters are chemicals that affect human or animal health by interfering with normal hormonal processes. They are mediated by hormone receptors, and so could

produce toxic effects at very low doses. Adherents to this hypothesis argue that applying standard risk assessments for toxic chemicals, which were developed mainly for carcinogens, does not adequately gauge the risk that endocrine disrupters pose to public health.

"The discussion at the first of these conferences was about whether 'low-dose effects' exist at all," says Hirozo Ueda, a director in the environmental health and safety division at the Japanese Environment Agency. "But now we are talking about which chemicals are associated with low-dose effects and which are not."

Researchers diverge sharply on the question of whether there is a threshold dose below which the chemicals can be regarded as safe. "There is no consensus yet that thresholds do not exist," says Robert



Kavlock, director of the reproductive toxicology division of the National Health and Environmental Effects Research Laboratory at the US Environmental Protection Agency.

Results from animal studies on suspect compounds, such as bisphenol A, a chemical widely used in the manufacture of plastics and glues, are inconclusive. "I am particularly perplexed by the data for bisphenol A, where some labs see it as a relatively potent oestrogen, and others fail to see much activity. This enigma needs to be understood at the mechanistic level," Kavlock says.

Pointing to data indicating that there is large genetic variability in the response to endocrine-disrupting chemicals, some scientists now argue that research in toxicogenetics, and novel tools such as gene-expression profiling and proteomics, may hold the key to understanding the effects of these chemicals in animals and humans.

"The evidence is that there can be as much as a 1,000-fold, or greater, range of responses to these chemicals in different strains of mice. The regulatory default assumption of a ten-fold correction or safety factor for genetic variability is completely out of touch with the data," says Frederick vom Saal, a professor of biology at the University of Missouri-Columbia.

"Toxicogenetics clearly has the potential to make a difference to how toxicology is conducted," says Kavlock. But he points out that there are still problems with the approach. "It remains to be shown that these new technologies can indeed provide tools for understanding the modes of action of endocrine-disrupting chemicals, and whether they can be used to monitor exposure signals in populations," he says.

Government regulation of endocrine disrupters is still years away. The US Environmental Protection Agency is undertaking a programme to validate test methods for the chemicals that will take several years to complete. No regulation can be implemented until then. And officials in Japan say that regulation there is at least ten years off.

Spanish leader lends ear to science

Xavier Bosch, Barcelona

Spanish scientists, whose discontent with the workings of the country's science and technology system has been growing, last week won an unexpected opportunity to take their concerns directly to the prime minister, José María Aznar.

Aznar met six leading scientists on 9 January at the presidential palace, along with his science adviser Pablo Vázquez and the heads of the Ministry of Science and Technology and the research arm of the Ministry of Health.

The prime minister told the delegation that the planned 2000–04 public funds for

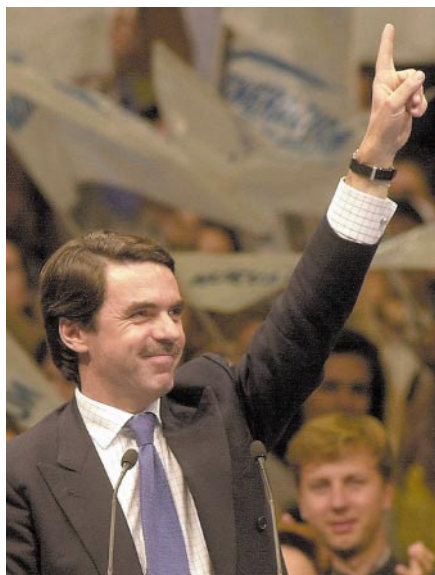
research were unlikely to be increased (*Nature* 402, 223; 1999). But, according to officials present, the meeting agreed on the acute need to prioritize areas of research more effectively, and to concentrate resources to build internationally competitive research teams such as that at the National Centre of Biotechnology in Madrid.

Other areas of agreement were the need to implement a continuous evaluation system for research, the need for new types of employment contract for researchers in the public sector, and the need for better links between the public and private sectors.

The scientists at the meeting included Mariano Barbacid, head of the National Centre for Cancer Research; Salvador Moncada, science adviser to the Institute of Cardiovascular Diseases and director of the Wolfson Institute for Biomedical Research at University College London; and Pere Puigdomènech, head of the Molecular Biology Institute in Barcelona. Juan Ramón Alaix, president of the Pharmacia drug company in Spain, represented the private sector.

The dialogue was very positive, says Barbacid, who is hoping it will increase the prime minister's support for Spain's scientific development.

Vázquez says that the meeting was Aznar's own initiative, and that the prime minister plans further meetings with other scientists every two or three months. "Aznar wants to obtain information directly from the horse's mouth concerning the problems faced by the scientific community," Vázquez says.



Well met: Spain's prime minister Aznar wants to address the science community's concerns.

CNN founder launches arms control initiative

Irwin Goodwin, Washington

Lately, international arms control has looked like a cause fast falling out of fashion. Memories of the cold war are fading, the population thinks about nuclear danger less, and, in the United States, the new administration of George W. Bush expresses more interest in missile defence systems than in arms control agreements.

So those who see arms control as a matter of critical importance have reacted with enthusiasm to last week's announcement by Ted Turner, the wealthy founder of the CNN television network, that he will sponsor a major new arms control institute to the tune of \$250 million over five years.

The non-profit Nuclear Threat Initiative will be devoted to reducing the threat of nuclear arms and other weapons of mass destruction, Turner says.

He convinced former Senator Sam Nunn to become vice-chairman of the organization. Nunn, a Democrat who represented Georgia, in turn enlisted two prominent Republican senators, Richard Lugar of Indiana and Pete Domenici of New Mexico, to serve on the initiative's board.

Turner says that the new organization's mission is "even more urgent now as the nuclear threat seems to have fallen off most people's radar screens since the cold war ended".

Nunn and Charles Curtis, a highly regarded former deputy secretary of energy who will be the organization's chief operating officer, have already met with defence ministers in Britain and France to win support for their work.

The new organization does not plan to duplicate existing government programmes, or provide grants directly to research in arms control. Instead, it will seek to raise public consciousness and understanding of the problems and to gain government support for arms control initiatives. "We intend to be a catalyst in education and to encourage

change in reducing pressure on the nuclear trigger," Curtis says.

One of the organization's priorities will be to support educational activities through universities, non-governmental groups and government agencies.

In an interview, Curtis emphasized the need to invigorate scholarship in arms control. "There has been a steady decline in arms control as a career path," he says. Stanford University in California and Harvard still teach arms control, but the Massachusetts Institute of Technology has abandoned its programme. ■



Prayers for peace: from left, Sam Nunn, Richard Lugar and Ted Turner at the organization's launch.

BSE fallout sends shock waves through Germany

Alison Abbott, Munich

Consumer panic has disrupted German meat markets after the discovery of a small number of bovine spongiform encephalopathy (BSE) cases in German-

born cattle. As the government attempts to get a grip on the crisis, the agriculture and health ministers have lost their jobs, and food-supply regulation and research are set for a major shake-up.

Andrea Fischer, the health minister and a member of the Green Party, resigned last week, following accusations that she mismanaged the recent crisis. She lost credibility after making confused statements over which types of sausages contain beef.

And Social Democrat Karl-Heinz Funke, the agriculture minister, resigned after criticism that his department had been slow to enforce a ban on the use of mechanically recovered meat in sausages.

The replacements for the two departed ministers will maintain the existing balance in Germany's governing coalition of Greens and Social Democrats. The new health minister is Ulla Schmidt, a Social Democrat, but part of her ministry's competence — consumer protection — will move to the agriculture ministry.

In an attempt to restore consumer confidence, the agriculture ministry is being

renamed the Ministry for Consumer Protection, Food Safety and Agriculture. Its new boss is Green Renate Künast, who has already responded to the BSE outbreak by announcing plans to encourage movement away from intensive agriculture and towards organic farming. The percentage of agricultural land given over to organic agriculture should increase from 3% to 10% in the next five years, Künast says.

Inclusion of consumer protection in the agriculture ministry is an attempt to temper influence of the agricultural lobby over the ministry. It is likely that the Federal Institute for Health-Related Consumer Protection and Veterinary Medicine, one of the health ministry's three scientific offices, will also move to the new agriculture ministry, where it will probably be reorganized.

Like some of its companion offices, the institute has been criticized as inefficient. A report by the Wissenschaftsrat, Germany's science council, suggesting better coordination between ministerial offices, is expected to be approved this week. ■



Love for wurst has given way to suspicion.

Fears arise that Bush may split science and technology office

Washington Scientific leaders in Washington are alarmed by unconfirmed reports that George W. Bush, the president-elect, plans to split the Office of Science and Technology Policy in the White House into separate offices dealing with science and technology.

George Trilling, the president of the American Physical Society (APS), is among those who have warned the Bush–Cheney transition team against splitting up the office, whose function is to coordinate science and technology policy across the government.

Any move to divide the office may be driven by a perceived need for a technology office and adviser to deal, primarily, with Internet- and telecommunications-related issues. APS newsletter *What's New* reported that two university presidents have already refused the resultant, diminished role of science adviser to Bush.

Scientific societies worry that a division would severely weaken the influence of the office, which also serves as the scientific community's voice in the White House.

Previous Republican administrations have fought with the office's predecessor organizations over the technical feasibility of

missile defence, which many scientists doubt. Bush has repeatedly stressed his intention to deploy a national missile defence system.

Chief scientist hopes to win Canadians' confidence

Montreal Canada has appointed a chief scientist to its federal health department for the first time. The move is a bid to restore public confidence following a scandal over HIV infection of the blood supply.

The chief scientist will be Kevin Keough, a biochemist who is currently vice-president of research at Memorial University in Newfoundland. His appointment takes effect in April, after which he plans to maintain his research group at the university.

Keough accepts that the department faces difficulty recruiting able scientists and says "we need to find mechanisms to allow people to express themselves" when working for the government.

Bioethics group looks at behaviour research

London Genetic tests for intelligence or antisocial behaviour may be some way off, but the Nuffield Council on Bioethics is already working on the issue. It has convened

a working party to examine the ethics of research into the genetics of human behaviour.

The panel is charged with examining what constitutes 'normal behaviour' before it considers the impact that an improved understanding of the genetics of behaviour will have on the future of education, insurance and the law. "The line between a complex disease and the spectrum of normal variation is a fluid one," says Martin Bobrow of the Cambridge Institute for Medical Research, a member of the working party.

The 12-member panel will begin consultations with the public in March and is expected to deliver its results in 2002.

♦ <http://www.nuffieldfoundation.org/bioethics>

CERN to lead work on data-sharing project

Munich The European Union has given nearly 10 million euros (US\$9 million) to a consortium of research organizations that is developing DataGrid, a project to help researchers share data over the Internet.

The project will be led by CERN, the European Laboratory for Particle Physics, near Geneva. One of its first challenges will be to handle the volume of data generated by the Large Hadron Collider, CERN's new accelerator, which is due to come online in 2005. But DataGrid will also facilitate the

handling of data in other scientific disciplines including astronomy and biology.

DataGrid is intended to link different computer resources across the world through a high-speed network, allowing databases of up to a petabyte in size to be shared. A petabyte is approximately the data content of a one-mile-high stack of CD-ROMs.

Institute gets second director in 30 years

Washington Paul Sieving, a professor of ophthalmic genetics at the University of Michigan, Ann Arbor, will become only the second director of the US National Eye Institute this spring, succeeding Carl Kupfer, who stepped down last summer after 30 years.

He will oversee a \$510 million budget, 300 staff on the National Institutes of Health campus in Bethesda, Maryland, and some 1,600 research grants and training awards.

El Niño devastated Indian coral reefs, study reveals

New Delhi Scientists in India have concluded that recent widespread damage to the country's three coral reefs was caused by warming during the 1997–98 El Niño event.

A study by the Centre for Ecological Research and Conservation in Mysore reveals

that bleaching occurred in 80% of the reefs on the Lakshadweep archipelago and in the Gulf of Mannar; more than 20% of their coral has died. Coral on the Gulf of Kutch reefs on the northwest coast suffered 11% damage.

The assessment says that El Niño caused the temperature in the Indian reef waters to increase by more than 3 °C. "Within three months ... between 80% and 90% of the coral was bleached or dead in the Lakshadweep and Mannar reefs," it says.

The survey was carried out in the summer of 1998, but subsequent observations suggest that the final damage was greater, with the reef reduced to 5% live cover in some places.

Gaur's death a setback for cloning hopes

Washington Last week's death of a cloned gaur (Indian bison) hit hopes that endangered species could survive with the help of cloning techniques that use surrogate mothers from related species. Noah, the baby gaur, died of a bacterial intestinal infection 48 hours after his birth to a normal cow in Illinois.

With only 30,000 remaining specimens, gaur — rare wild oxen native to India and Southeast Asia — are threatened with extinction. Scientists at Advanced Cell Technology (ACT) based in Worcester,



No ark for Noah: the baby gaur's death confirms problems with the cloning of endangered species.

Massachusetts, used the 'Dolly' technique (see *Nature* 385, 810–813; 1997) to produce clones from a male gaur that had died eight years ago. Eight out of 42 cows implanted with gaur embryos became pregnant, but Noah's mother was the only one to give birth.

Noah's death offers some backing to sceptics who questioned the value of cloning techniques for assisting the survival of species. But Robert Lanza, the head biologist at ACT, is not deterred. The Spanish government last year approved his plan to produce clones of the recently extinct Pyrenees bucardo goat, using normal goats to carry the embryos.

Polio's last stand

Does an outbreak of poliomyelitis in the Caribbean caused by a mutated vaccine mean that plans to complete the disease's eradication must be reworked? Tom Clarke considers the evidence.

Hot spot: the Dominican Republic recently played host to an outbreak of polio caused by rogue viruses from the oral polio vaccine.

Initially, pesticide poisoning was the suspected cause. But epidemiologists investigating reports last summer that children in the Constanza region of the Dominican Republic were suffering from a mysterious paralysis discovered something a little more sinister. The researchers, from the Pan American Health Organization (PAHO), found that the children's lifeless limbs bore all the hallmarks of polio — a disease that the PAHO had certified as being eradicated from the Americas in 1994.

If the epidemiologists were surprised at polio's apparent return to the Caribbean, it was nothing compared with what was to follow. In late October, laboratory analysis confirmed that the virus infecting the Dominican children was polio, but it was not a wild strain. Instead, it was a mutant form of a live virus from the oral polio vaccine (OPV), used to stamp out the wild strains in the first place¹. This virus, deliberately weakened for use in the vaccine, had regained the ability to cause an outbreak of disease — something never previously seen in almost four decades of field experience.

This unexpected discovery is causing soul-searching among public-health officials working towards the global eradication of polio. In the same month that the Caribbean polio cases were confirmed, the World Health Organization (WHO) had declared the entire western hemisphere free of wild poliovirus. The OPV's unquestioned success had allowed the WHO to begin planning its polio 'end-game' — the final stage in its global polio eradication initiative.

The idea was carefully to withdraw OPV from use. Once the weakened viruses in the vaccine died out, poliomyelitis would join smallpox on the roster of conquered infectious diseases. But if the viruses from the OPV circulating in the environment can revert to a disease-causing form, this strategy might be dangerous. Indeed, some experts are now wondering whether it will be necessary to use another vaccine, incapable of such reversion, to eradicate the OPV viruses.

In the 1960s, when doctors first began dropping the OPV into the mouths of children, they knew that a vaccine made from live, weakened polioviruses could survive in the environment and spread from person to person. Indeed, this attribute is one reason why it has been so effective. Children given



Into the mouths of babes: the oral polio vaccine has been key in the fight to eradicate the disease.

the OPV continue to excrete the vaccine's viruses into the environment for some time after they are vaccinated. By infecting people who are later exposed to these excreted viruses, the vaccine confers some immunity to individuals missed by vaccination efforts — a real benefit in developing countries, where vaccination coverage can be poor.

Mutant manoeuvres

Cases of polio caused by the vaccine are not unknown. About one in every 750,000 people vaccinated with the OPV becomes ill — mostly people with weakened immune systems. Virologists also know that, under selective pressure in the harsh environment of the gut, viruses in the OPV can mutate into more virulent strains². But in nearly 40 years of experience, these excreted mutant viruses had never been shown to start an outbreak of polio.

At the latest count, seven cases of vaccine-derived polio have been confirmed in the Caribbean — six in the Dominican Republic and one in neighbouring Haiti. An eighth case, from the Dominican Republic, is awaiting laboratory confirmation at the US Centers for Disease Control and Prevention (CDC) in

AFP

CORBIS

Atlanta. After analysing the gene sequences of the viruses found in the victims, PAHO epidemiologists are certain that all but one of the cases in the Dominican Republic are epidemiologically linked. This means that not only has one of the OPV viruses become virulent, but that it has recovered the ability to spread.

This disturbing outbreak is a dark cloud on an otherwise bright horizon. Since the WHO's global polio eradication initiative was launched in 1988, the number of cases of polio worldwide has fallen by over 99% — from 350,000 in that year to about 2,000 last year. The WHO officially declares a region polio-free when no cases of polio caused by wild virus strains have been recorded for three years. The Western Pacific joined the Americas in being certified polio-free in October 2000, completing the disease's eradication from the western hemisphere. The WHO European region, which includes the former Soviet Union, has not seen a new case of polio since the end of 1998. And although countries in Africa and Asia still play host to annual epidemics, the WHO expects to rid the world of polio by 2005. "We're very, very close," says Bruce Aylward, who coordinates the eradication initiative from the WHO's headquarters in Geneva.

Triple whammy

The WHO's success with polio is largely attributable to the OPV, developed in the 1950s by Albert Sabin, then at the University of Cincinnati College of Medicine in Ohio. Sabin's OPV replaced Jonas Salk's famous vaccine, which was made using dead poliovirus. Sabin patiently created the OPV by passing each of the three types of naturally occurring poliovirus through different non-host tissues under various culture condi-

tions. After each passage, he screened the surviving viruses for their ability to cause disease in monkeys. By selecting the weakest strains of each virus and continuing the passing regime, Sabin managed to produce a live vaccine containing weakened viruses that would confer immunity to all three types of polio.

The OPV stimulates the production of antibodies in the blood, protecting the nervous system from the neurotoxic effects of the virus should infection occur. But because the vaccine is given orally, it also produces a local immune response in the intestine's mucosal membranes, causing the production of both antibodies and T cells, which destroy infected cells. This attacks the virus as soon as it gains an initial foothold in the gut, rapidly shutting down person-to-person transmission of wild polio strains if enough of the population is vaccinated. "In terms of eradicating wild-type polio viruses, there's no question that the OPV works," says David Wood, a polio expert at Britain's National Institute for Biological Standards and Control in Potters Bar, near London.

The potential for OPV viruses to revert to a dangerous form has always been there. But to the encouragement of WHO officials, this reversion did not seem to be happening in the real world. Studies of children in Cuba, who all receive the vaccine at the same time of year, indicated that although the OPV is excreted and circulates in the population, it disappears within three months³.

Analyses carried out by Roland Sutter, chief of the polio eradication branch at the CDC, suggested that OPV viruses are slow to mutate. In an unpublished study of isolates taken from vaccinated populations across the Americas between 1995 and 1997, Sutter and his colleagues have found that gene



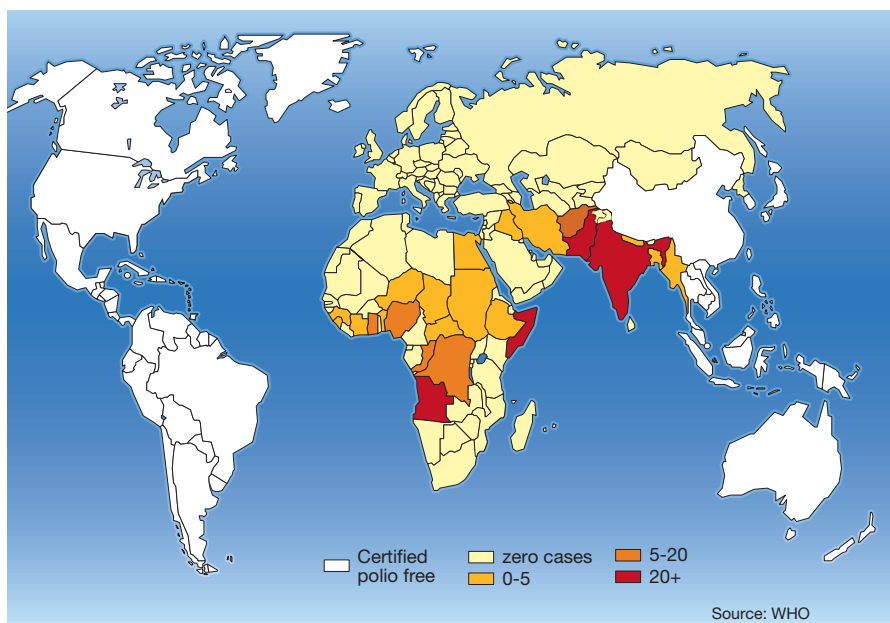
Devastating results: the neurotoxic effects of polio can leave its victims paralysed.

sequences from the viruses were 99.5% identical to those from a sample of the OPV straight from the shelf. "The isolates suggested we didn't have a problem," says Sutter.

But the OPV-derived polio outbreak in the Caribbean throws that comforting view into question. If problems with the OPV were ever going to emerge, the island of Hispaniola, shared by the Dominican Republic and Haiti, was a likely venue. It is an area where there has been no wild polio for several years, so natural immunity is low. In addition, continued vaccination coverage was poor. In the Constanza region, only about 20% of children in 2000 had received one of the three doses of OPV needed to give adequate immunity. "Essentially it was a totally susceptible population," says Ciro de Quadros of the PAHO, who has led the campaign against polio in the Americas. Unpublished analysis of the gene sequence of the rogue OPV virus carried out by the CDC suggests that it had been replicating for as long as two years — long enough to turn nasty.

Back to the drawing board?

Now the WHO faces some difficult questions in planning its polio eradication end-game. Optimists had hoped that this would comprise a relatively simple — albeit difficult to organize — *coup de grâce*: a coordinated withdrawal of OPV on the same day worldwide. Provided all populations had been adequately vaccinated, the OPV viruses should just disappear from the environment as they have been seen to do in Cuba. Now the WHO cannot be sure that this will happen.



Close to extinction: the incidence of polio caused by the wild virus in 2000 shows that the programme to control the disease is close to success, but the figures do not include vaccine-induced cases.

▶ “Not only does the OPV revert to neurovirulence and paralyse people, but it also looks as if it is really transmissible,” says Sutter.

First on the list of priorities is to figure out just how likely another OPV-derived outbreak will be. “My assumption is that this is an unusual event,” says Donald Henderson of Johns Hopkins University in Baltimore, who led the WHO’s successful effort to eradicate smallpox. “It can’t be occurring with great frequency otherwise we would have seen it a long time ago.” But polio experts freely admit that, because they did not suspect the OPV in the past, they may have missed previous vaccine-related outbreaks. In fact, a retrospective analysis of isolates from sporadic polio cases in Egypt between 1989 and 1998 indicates that at least some of the cases there may have been due to vaccine-derived polio⁴. This was not noticed at the time because Egypt is still host to wild virus strains.

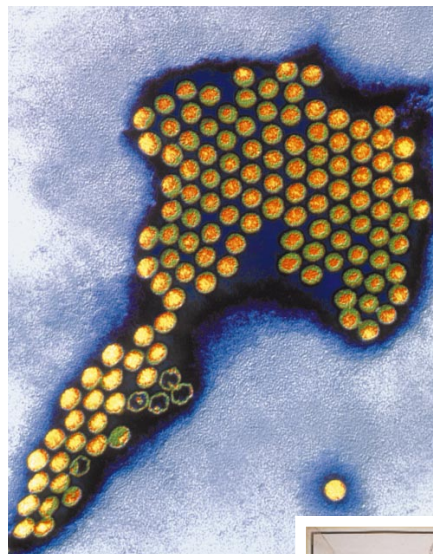
One key step towards determining the likelihood of future vaccine-derived outbreaks is establishing how thin the vaccination coverage must be before the frequency of mutant viruses begins to rise from the low levels seen in the Americas by Sutter. The cases in Hispaniola and Egypt occurred in areas where vaccination coverage was very poor, meaning that a greater number of people with no immunity were exposed to potentially virulent OPV viruses. But how patchy a vaccination effort can be before a crisis occurs is not known.

Gaining a better understanding of the genetic determinants of viral transmissibility will also be important. Although the basis for neurovirulence in polio is fairly well understood — for example, point mutations in the gene for the virus’ protein coat alter its ability to paralyse its victims⁵ — Wood describes the molecular basis of polio transmission as “one of the key unknowns”. If this information could be gleaned, he points out, it would be possible to identify quickly mutants that have the potential to cause outbreaks.

There is also a need to determine factors that might allow the OPV to persist in the environment for longer than anticipated. For instance, recent research has shown that people with severely impaired immune systems can remain host to the OPV viruses for many years — possibly serving as reservoirs once polio has been eradicated⁶. Although initial investigations have found no evidence for HIV/AIDS patients harbouring the OPV viruses, it could be an important factor for polio eradication in countries with high rates of HIV and low polio vaccination coverage⁷.

But the problem is that, as the WHO’s eradication campaign proceeded, research on polio slipped down the list of biomedical priorities. “The funding agencies are turning round to people and saying ‘you should be moving on to something else,’” says Wood.

It is too early to tell whether the Caribbean outbreak will lead to renewed funding for polio research. But officials with the WHO



Wipe out: after a twelve-year campaign, the poliovirus (above) is set to follow smallpox as a conquered infectious agent.

and the PAHO argue that Hispaniola provides an excellent arena for answering some of the important questions. “The investigation of this is hugely important,” says Aylward. “It will help us figure out how to track the virus epidemiologically and virologically.” Indeed, analysis of the RNA sequences of OPV isolates from Hispaniola are already allowing the CDC to track the course of the rogue virus’ spread and may help reveal how it recovered its ability to cause disease outbreaks in the first place.

Under observation

Although a thorough evaluation of what went wrong on Hispaniola will yield information crucial to the WHO’s polio end-game, other obstacles remain. According to Aylward, surveillance for vaccine-derived polio must be improved. Unlike the disfiguring rash of smallpox, polio’s symptoms are tough to spot and must be confirmed in a well-equipped laboratory. This makes it difficult to identify outbreaks in remote parts of the world where there are few trained health workers.

More fundamentally, if it becomes clear that OPV-derived outbreaks are more likely than was thought, another vaccine may be needed to maintain people’s immunity to polio while the OPV strains die out. Right now, the only available candidate is the Salk inactivated polio vaccine (IPV). This requires only one dose, confers immunity to all types of polio, and, because it is made from dead virus, cannot cause disease.

But the IPV has some serious drawbacks for use in developing countries. First, it does not trigger mucosal immunity in the gut and

so does not stop transmission. Second, it must be injected, meaning that it must be administered by trained health workers, rather than unqualified volunteers. And there is currently no industrial capacity to produce the vaccine on the scale required. “The cost is going to be mind-boggling,” says Henderson.

Another solution would be a totally new vaccine. Approaches based on subunits of the virus’ protein coat, which work well against many diseases, do not look promising for polio. But it is now possible to use the techniques of genetic engineering to create live vaccines that lack genes for neurovirulence — which should be very unlikely to revert to a virulent form. But new vaccines take time and money to produce, and the WHO opted early in the eradication programme not to develop

improved vaccines — a decision described in 1999 by Henderson, writing in the journal *Vaccine*⁸, as “an extraordinary act of ignorance”. Now, with polio on its way out, few vaccine manufacturers are likely to be interested in bringing a new vaccine to market. “This is going to be at best a vaccine with a very limited lifespan,” says Aylward.

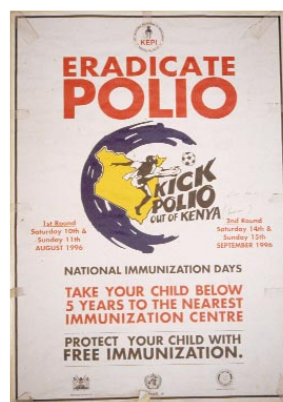
The good news is that on Hispaniola tests of all subsequent suspected cases of polio have proved negative — indicating that the brief epidemic is over. The authorities in Haiti and the Dominican Republic are determined not to get caught out again. “Both countries are very worried and are responding very seriously to the situation,” says de Quadros. Last month the Dominican Republic held a massive campaign to vaccinate 1.2 million children with the OPV in three days. And earlier this month, Haiti began a similar effort, now being overseen by de Quadros.

But the WHO’s polio experts face an anxious future, as they wait for the results of investigations into the Hispaniola outbreak. At best, it may come to be seen as a regrettable case of inadequate vaccination and surveillance, reinforcing the need to follow the WHO’s recommended regime. At worst, it may mean that plans for the polio end-game need to be redrawn. “Clearly this is raising a red flag,” says Sutter. “Whether it’s a small one or a big one is something we need to worry about.” ■

Tom Clarke is a science writer based in London.

1. *MMWR Morb. Mortal. Wkly Rep.* **49**, 1094 (2000).
2. Abraham, R., Minor, P., Dunn, G., Modlin, J. F. & Ogra, P. L., *J. Infect. Dis.* **168**, 1105–1109 (1993).
3. Mas Lago, P. et al. *Bull. World Health Organ.* **72**, 221–225 (1994).
4. Naguib, T., Yang, S. J., Pallansch, M. & Kew, O. *MMWR Morb. Mortal. Wkly Rep.* (in the press).
5. Evans, D. M. A. et al. *Nature* **314**, 548–550 (1985).
6. WHO Scientific Group *Clin. Exp. Immunol.* **109** (Suppl. 1), 1–28 (1997).
7. Fine, P. E. M. & Carniero, I. A. M. *Am. J. Epidemiol.* **150**, 1001–1021 (1999).
8. Henderson, D. A. *Vaccine* **17**, S53–S55 (1999).

▶ <http://www.polioeradication.org>



Funding assured for India's international biotechnology centre ...

Sir— We are writing to you in our capacity as president of the board of governors of the International Centre for Genetic Engineering and Biotechnology (ICGEB), and director general of the ICGEB, respectively. We are utterly surprised by the content and the tone of your Editorial and News item about this centre (*Nature* **408**, 121 & 127; 2000).

Last November a meeting of the board of governors, representing the 43 countries involved, very positively appraised the activity of both the Italian and the Indian components of the ICGEB and confirmed the board's total confidence in the ICGEB management, including Professor V. S. Chauhan, director of the New Delhi component. This positive evaluation was also based on a glowing report by the ICGEB council of scientific advisers — an independent body of 15 distinguished scientists.

The board was totally satisfied that the turnover of personnel is at a normal level for a research institution and that the hiring of all personnel is done in full transparency, without any external pressure, political or otherwise, but purely on professional grounds. Furthermore, the morale of the scientists working at the ICGEB is quite good, as witnessed by the accompanying letter that has been sent to you by the New Delhi scientific staff.

Their morale has been further heightened by the outcome of the board's meeting: the government of Italy has just passed a law to triple its annual contribution to the centre (hardly a sign of "donor fatigue") and the government of India has promised to do likewise. All the other governors present at the meeting have assured that their countries will fulfil their financial commitments to the ICGEB. Furthermore, grants to ICGEB scientists from several major international foundations and agencies are steadily on the increase.

Adolfo R. Taylhardat, Arturo Falaschi

International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34012 Trieste, Italy

... and senior staff say there's no problem

Sir— The description of the working conditions and scientific atmosphere at the International Centre for Genetic Engineering and Biotechnology (ICGEB), in your News report (*Nature* **408**, 127;

2000) comes as a surprise to us. We represent 90% of the senior investigators working at the ICGEB, New Delhi, and we do not agree that there have been "deteriorating working conditions" at our centre. If *Nature* had solicited views from a wide range of scientists working at the ICGEB you would have got a completely different view of the working conditions and atmosphere here.

The healthy scientific atmosphere at the ICGEB, New Delhi, is reflected in the scientific output of the centre. In addition to a growing list of publications in international scientific journals, ICGEB investigators have been successful in attracting grant support from a wide range of international funding agencies including the Wellcome Trust, Howard Hughes Medical Institute, the World Health Organization, the European Commission, the Rockefeller Foundation, the (US) National Institutes of Health and the Human Frontiers Science Program in recent years.

Moreover, the activities of the ICGEB are reviewed annually by an independent council of scientific advisers (CSA), which includes two Nobel laureates. The two CSA members quoted by your correspondent had only positive things to say about the scientific work of the ICGEB, New Delhi. We believe that the quality of our scientific research is a direct reflection of the healthy working atmosphere prevailing here.

S. K. Sopory

Other signatories to this letter:

V. Siva Reddy, Raj Bhatnagar, Sunil Kumar Mukherjee, Nirupama Bhatnagar, Kanury V. S. Rao, Suresh Nair, N. Tuteja, Neeti Sanan Mishra, Nirupam Roy Choudhury, M. K. Reddy, Raman Rajagopal, Krishnamurthy Natarajan, Suman Dhawan, Chetan Chitnis, Pawan Malhotra, Sadhu Leelaavathi, Ashima Kushwaha, Sunil K. Lal, Deepak Sehgal, S. Swaminathan, N. Khanna, Shahid Jameel, Anand Ranganathan
International Centre for Genetic Engineering and Biotechnology, Aruna Asaf Ali Marg, 110 067 New Delhi, India

Doing our best for returners to Spain

Sir— The letter signed by Pau Ferrer *et al.* (*Nature* **407**, 941; 2000) accurately describes the severe problems that young researchers face in rejoining the Spanish research system after several years abroad. Although I fully share their frustration and concern, the letter contains some statements that need clarification.

The writers mention a "lecturing policy document" produced by the Universitat Autònoma de Barcelona (UAB). Our policy has been to create transitory five-year posts for people who already had a very

precarious teaching contract with the UAB. The posts are occupied by PhDs, some of them with research experience abroad. Such positions are equivalent to teaching assistantships in the US academic system (not to assistant professorships, as the letter says). After the fifth year, there will be tenured positions by public and open competition, and any citizen with a PhD is welcome to apply.

Academic rights within the UAB were established through our democratic bylaws 15 years ago; they are our 'constitution'. Unfortunately we could not predict, so long ago, that the Ministry of Education would select and fund reincorporated scientists. This is why they are not now represented on the university's elected body. This issue, however, may be revised in the near future to allow for this and other unforeseeable situations. In other important aspects of academic life, people in these situations face no discrimination, to the best of my knowledge.

The scientific community considers it an absolute priority for Spain to improve its efforts in R&D. Let me emphasize that we consider young researchers with experience in good universities and prestigious research centres to be essential for the future of our society.

Carme Picallo

Universitat Autònoma de Barcelona, Edifici A—Rectorat, 08193 Bellaterra (Barcelona), Spain

Difficulty in reconciling global-warming data

Sir— According to your Special Feature on global warming (*Nature* **408**, 896; 2000), a scientific report issued by the US National Academy of Sciences¹ concluded that the discrepancies between temperature measurements made at the surface (showing a warming trend since 1979) and from satellites (showing little if any tropospheric trend) "were now reconciled and that they pointed to a warming planet".

To be precise, the NAS panel, which included some of the severest critics of the satellite data, could not reconcile the disparity between the two trends. In fact, the disparity is opposite in direction to expectations based on climate models: climate models all predict a stronger warming trend for the mid-troposphere than for the surface.

S. Fred Singer

Science & Environmental Policy Project, 1600 South Eads Street, Suite #712-S, Arlington, Virginia 22202-2907, USA

1. National Research Council/National Academy of Sciences
Reconciling Observations of Global Temperature Change
(National Academy Press, Washington DC, 2000).

How Maxwell made his mark

His findings, taken for granted now, were fiercely contested during his life.

Electrodynamics from Ampère to Einstein

By Olivier Darrigol
Oxford University Press: 2000. 532pp.
£75, \$130

Scott Walter

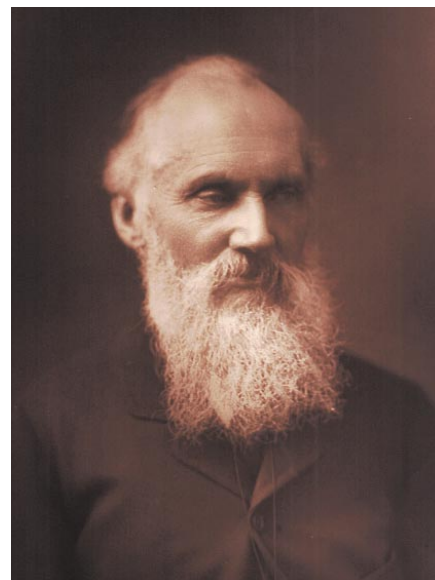
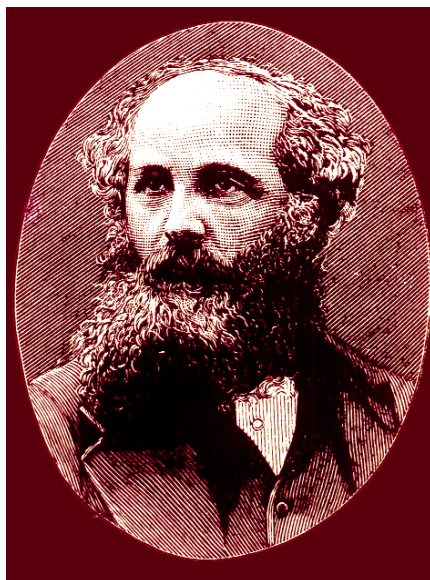
Among the enduring legacies of nineteenth-century science, James Clerk Maxwell's equations of electrodynamics have long held a preferential place in the hearts of physicists. One of today's more outspoken physicists, Steven Weinberg, has argued that the equations constitute a noncontingent fact, without which contemporary physics would be unimaginable. But, little more than a century ago, other, non-Maxwellian, futures were still imagined, not only by poets and science-fiction writers, but by the likes of Kelvin and Helmholtz, for whom all known electrodynamic effects were the result of forces acting at a distance, in contrast to Maxwell's theory of the electromagnetic field.

In his newest book, Olivier Darrigol shows how, at crucial junctures in the nineteenth century, such leading physicists held conceptually incompatible views of the nature of electricity and magnetism and, by extension, of the future direction of physics. With time, physicists initially sceptical of Maxwell's views came to embrace field theory, following the example set by Helmholtz, who conceded Maxwell's superiority over his own account in the mid-1870s. Kelvin, on the other hand, never gave in, not even after Hertz's ingenious demonstration of electromagnetic wave propagation in air.

During the past 20 years, historical investigations of the origins, discovery and reception of Maxwell's theory have transformed our image of a moribund Victorian science, with its bizarre mechanical models built to satisfy what Kelvin referred to as "fits of ether dipsomania", into one of high-stakes intellectual adventure, pursued by physicists in scientific centres across Europe.

Darrigol's book is the first to draw on this work on a large scale, and is the first comprehensive history of electrodynamics since E. T. Whittaker's *A History of the Theories of Æther and Electricity* (1910). It covers the period from 1820 to 1905, and stands as a chronological sequel to John Heilbron's *Electricity in the Seventeenth and Eighteenth Centuries* (1979). Major developments and turning points discussed include Ampère's law, Faraday's notions of charge and current, the emergence of new quantitative methods in Germany, Maxwell's equations, Hertz's experiments, and the advent of the electron.

Each of these topics has been addressed in



Maxwell (left) and Kelvin: Kelvin held out against Maxwell's ideas to the end.

monographs, but it is to Darrigol's credit not only to have brought this research together in an effective narrative but also to have filled in the gaps. For example, Darrigol discusses Helmholtz's investigations of the *RL* and *RLC* circuits used to model muscle contraction, because they effectively guided the latter's transition from physiology to physics. Informed by manuscripts recently discovered in Berlin, the author's analysis of Helmholtz's theory of electrodynamics holds particular interest. It shows how Helmholtz, after reinterpreting Maxwell's theory in terms of electric actions at a distance and an infinitely polarizable vacuum, convinced himself that the latter theory represented the only viable alternative for future research. This is only one example among many of an electrodynamicist making the 'right' choice for the 'wrong' reason.

Darrigol's broad overview of leading ideas of the time and their relationship to one another gives new insights into the emergence and evolution of theoretical and experimental research traditions. At the same time, it reveals remarkably different interpretations of Maxwell's equations by physicists in Britain and on the Continent. Most notably, the Maxwellian notions of electric displacement and current were misunderstood by continental physicists, including Hertz, who, after failing to make sense of Maxwell's potentials and hypothetical fluids, famously concluded that Maxwell's theory is Maxwell's system of equations. Packed in this epigram was a new theoretical outlook, according to which physical reality was best modelled by differ-

ential equations, and exemplified by Hertz's purely mathematical theory of the electromagnetic field.

For Darrigol, the historical unity of electrodynamics derives from a chain of ideas and events running from Ampère to Einstein, the links of which he patiently lays out for the reader. Surprisingly, only two of nine chapters have much mathematics: one is on Maxwell's theory, the other on electron theory. Anyone referring to the original texts in these areas is faced with a bewildering array of terms and symbols, along with some strange mathematics. Nineteenth-century physicists, Darrigol observes, were often just as baffled, but to reduce reader discomfort he imposes standard units and notation, while noting the resulting anachronisms. Even with familiar notation, of course, many exercises are left to the reader, but, for once, there is no compromise with substantial facts or issues. Derivations and extended technical explanations are relegated to the appendix, but technical or not, opinions are well documented, with appendices, bibliography and index accounting for nearly a quarter of the volume.

One of the recurring themes of Darrigol's history concerns the close intertwining of theory and experiment in nineteenth-century electrodynamics, which is distinguished by the fact that all leading theorists were active in the laboratory. Observing how individual electrodynamicists coordinated theoretical and experimental activity, Darrigol presents evidence that the same methodological principles were applied to both theory and experiment. In this way, Darrigol explains the

profound harmony between the theoretical and experimental practice of Ampère, Faraday, F. Neumann, W. Weber, Kelvin, Maxwell, Helmholtz and Hertz.

Darrigol's guided tour of the "lofty summits of the history of electrodynamics" will appeal to historians and philosophers of science, as well as to physicists, mathematicians, and engineers interested in the origins and evolution of field theory. Regardless of how one may feel about the chances for success of non-Maxwellian alternatives a century ago, Darrigol's informed analysis of the evolution of electromagnetic theory and experiment effectively illustrates the subtle ways by which Maxwell's equations came to shape visions of the future. ■

Scott Walter is in the Department of Philosophy, Archives Henri Poincaré, Université Nancy 2, 23 Bd Albert 1er, 54015 Nancy Cédex, France.

Not so crazy after all

Strong Imagination: Madness, Creativity and Human Nature
by Daniel Nettle

Oxford University Press: 2001. 192 pp. £16.99

Dylan Evans

"I have long had a suspicion," wrote the great Victorian psychiatrist Henry Maudsley in 1871, "that mankind is indebted for much of its individuality and for certain forms of genius to individuals [with] some predisposition to insanity. They have often taken up the by-paths of thought, which have been overlooked by more stable intellects." In *Strong Imagination*, Daniel Nettle takes up Maudsley's suspicion and runs with it. He argues that the genes that predispose people to schizophrenia and manic depression have been maintained in the gene pool by natural selection because of their beneficial effects in enhancing creativity.

The idea that madness and creative genius are but two sides of the same coin has a long and distinguished pedigree, originating long before Maudsley. In the past few decades, the idea has been subjected to ingenious statistical tests by Kay Redfield Jamison, Arnold Ludwig, Nancy Andreasen, Felix Post and others. Nettle reviews this literature, and suggests a couple of reasons why a little dose of madness might be a good thing for a creative artist; mania provides the energy and drive necessary for sustained lonely work, and schizotypy favours divergent thinking. Add to this the idea that creativity leads to greater reproductive success, and the result is an adaptationist account of schizophrenia. The account is adaptationist, not because it requires that schizophrenics have more babies — Nettle rightly draws back from this claim — but because it implies that the genes that predispose towards schizo-



Sylvia Plath: creative and depressed — but why?

phrenia have remained in the gene pool because they also enhance creativity.

From the chatty style, and the fact that even such basic biological entities as neurotransmitters and recessive genes are explained in simple terms, it is clear that the book is aimed at the general reader with very little knowledge of psychiatry or evolutionary theory. Such readers will no doubt learn a lot from the book, as it covers an impressive range of material, from the history of ideas about mental illness in the twentieth century and the basic principles of neurochemistry, to the phenomenology of psychosis and the theory of sexual selection. They may even find themselves persuaded by Nettle's thesis. Those already familiar with the literature, however, will find the book less convincing.

Nettle frequently claims to have "argued" for some claim or even "demonstrated" it, when in fact he has merely proposed or assumed it. The result is a text that may please the converted but will not persuade the sceptic. This is perhaps most obvious when Nettle discusses the supposed evolutionary advantages for psychological traits such as low/high mood and creative flair. "How," he asks at one point, "can low mood be adaptive, given that in all primate groups, status is positively related to reproductive success, and low mood makes us drop in status?" One obvious answer is that low mood might not be adaptive at all, but Nettle doesn't even consider this possibility. This is just the kind of approach that has got adaptationism a bad name, and which more cautious adaptationists such as George Williams have repeatedly criticized.

Nettle's grasp of evolutionary theory is much weaker than his grasp of psychiatry, which is generally sound. Reflecting this imbalance, the citation of authorities is also very uneven: in the chapters about psychosis

all the usual suspects are there, but when it comes to evolutionary psychology we find a very skewed sample. On the one hand, the major figures in the field are notable by their absence. The Machiavellian intelligence hypothesis is discussed (although not named), but there is no mention of Nicholas Humphrey, who first proposed it, and Robin Dunbar, who has done much to test this hypothesis, only gets a brief nod in the acknowledgements (Richard Byrne and Andrew Whiten don't even get that). Likewise, Nettle advances something suspiciously like Randolph Nesse's propitiousness hypothesis of mood without mentioning Nesse's name at all.

On the other hand, the authorities who are mentioned are among the least trustworthy, and several ideas are misattributed. Nettle seems particularly impressed by Anthony Stevens and John Price, although their book, *Evolutionary Psychiatry* (Routledge), is among the more dubious contributions to the field, and he credits them with the idea that the modern world has changed too fast for the human mind, which is still adapted to the pleistocene. Now, Stevens and Price may well have baptized this hypothesis with a particularly catchy name — they call it the 'genome lag' hypothesis — but it is certainly not their invention. It is, in fact, one of the staple ideas in evolutionary psychology, and goes back at least as far as John Bowlby.

The one evolutionary psychologist who is both a major player in the field and cited frequently in the book is Geoffrey Miller. Miller's view that sexual selection has played a much greater role than natural selection in shaping the most distinctively human aspects of our minds is linked suggestively with the view that the genes for madness are also genes for creativity. Indeed, the main virtue of Nettle's book is that it brings these two hypotheses together in a single work for the first time. This is a task that needed to be done, and we must be grateful to Nettle for undertaking it.

Despite the occasional stylistic infelicity and the overuse of the first person singular, Nettle writes well. He enlivens the scientific data with fascinating clinical vignettes, anthropological observations, and well-chosen quotations from Shakespeare. The book may not contain anything strikingly new or original, but it does constitute a readable and up-to-date review of a very large body of literature on a fascinating subject. The lack of novelty may seem odd in a book about creativity, but perhaps this is no bad thing. If Nettle is right, novelty is often purchased at the price of delusion. It is reassuring, then, to read that one of the studies cited in the book shows that, out of a wide range of "eminent people", scientists had one of the lowest lifetime rates of mental disorder. ■

Dylan Evans is in the Department of Philosophy, King's College London, 160 The Strand, London WC2R 2LS, UK.

A book that rocked the Victorian world

Victorian Sensation: The Extraordinary Publication, Reception and Secret Authorship of *Vestiges of the Natural History of Creation*

by James A. Secord
University of Chicago Press: 2001. 581 pp.
\$35, £22.50

David Oldroyd

Traditionally, historians of science have studied what scientists have thought and how they generated ideas, as well as the social, economic and political implications of science. Among geo-historians, controversies have attracted particular attention: in his *Controversy in Victorian Geology* (Princeton, 1986), James Secord wrote the definitive account of the contest between Adam Sedgwick and Roderick Murchison about the Cambrian–Silurian boundary. Now Secord, American by birth but long immersed in British culture — he is Reader in History and Philosophy of Science at Cambridge — re-orientates the task of science historians in what will become another definitive book. He asks not so much how Robert Chambers's *Vestiges* was produced as how it was received, by whom and where. It is a grand but daunting undertaking.

Secord makes the remarkable claim that his book is "the most comprehensive analysis of the reading of any book other than the Bible ever undertaken". I'm not qualified to judge the truth of this assertion, but certainly it is plausible. Immense pains have been taken to set the scene for the publication of the book and to describe, through exhaustive historical research, how it was distributed and sold, and most particularly how it was read — by whom, why, and with what effects on different readers. Thereby the author constructs a remarkably intricate picture of major features of nineteenth-century British culture. All this is undertaken in a book which, despite its long and complex narrative and argument, is both coherent and agreeably written, besides being adorned with 155 illustrations, mostly unfamiliar.

Vestiges (1844), of course, wasn't just any old book. It was a sensation, and a significant piece of the darwinian jigsaw. The fact that it was published anonymously made it special, although anonymity was commoner then than now: reviews were often unsigned and authors (women particularly) often sheltered behind pseudonyms. Chambers opted for anonymity because his cosmic evolutionism — with spontaneous generation and a gradual 'unfolding' of 'higher' life forms in a 'gestatory' process — was incompatible with the deemed truths of revealed religion and potentially

The treasures beneath our feet



Excellent pictures and straightforward, informative text make Ron Vernon's *Beneath our Feet: The Rocks of Planet Earth* (Cambridge

University Press, £18.95) almost a coffee-table book for the scientifically literate non-geologist — or anyone who's always curious.

subversive of social order. Hence the Edinburgh publisher and amateur scientist wore the cloak of anonymity to protect his business interests, using another firm to issue his book.

Some of Chambers's ideas were implausible, but they included a disturbing actuality: the 'parallelism' between embryonic development, the stratigraphic record, and the principal features of animal classification. These analogies suggested causal connection. This hinted that revealed religion (as then conceived) was deficient — which implied that the social 'cement' might crack if something was not done, and quickly.

The work's anonymity unsettled its reviewers. Were they writing about Charles Lyell, Ada Lovelace, the phrenologist George Combe, the Prince Consort ...? Caution was needed. Adam Sedgwick, however, secure in his Trinity sanctuary, was unrestrained. Writing anonymously in the *Edinburgh Review* — but in his case making sure that his authorship was known — he averred that *Vestiges* had "annulled all distinction between physical and moral". Pulling rank, he asserted that "[n]o man living, who has not partaken of this kind of [scientific] labour, or ... has not thoroughly mastered the knowledge put before his senses by the labours of other men, has any right to toss out his fantastical crudities before the public, and give himself the airs of a legislator over the material world". Sedgwick had some reason to complain. He had toiled over many mountains, done much original research, mastered many books, fought battles at the Geological Society. His anonymous opponent evidently had not. But Sedgwick was also worried. If *Vestiges* were correct, his position might be questioned. The comforts of (the) Trinity, or his clerical position at Norwich, might have no moral warrant.

These matters are relatively well known.

But Secord pursues them further. He has examined Sedgwick's annotated copy of *Vestiges*, showing how he scrawled all over it. He ascertains which passages attracted Sedgwick's attention, and how he responded to them. Similarly, Secord shows how the Halifax autodidact, Thomas Hirst, doggedly wrote out extracts from *Vestiges*, assimilating them with, as Secord puts it, "extraordinary intensity". "Almost no one," he continues, "reads like this any more." True. Books were scarcer items then than now, and everything possible had to be wrung from them by the earnest self-improver.

Of course, Darwin's shadow falls over this. But in his epilogue, Secord argues that it was not just Darwin's work that changed everything, creating a new world-view. For example, although *The Origin of Species* appeared in 1859, it was selling fewer copies than *Vestiges* until the 1880s. Focusing all attention on Darwin obscures rather than clarifies the past. So by concentrating on the production, distribution and reception of Chambers's book and the "network of relations that make up the larger picture", rather than the factors that led him to arrive at his ideas, Secord aims to offer "an experiment in a different kind of history".

Secord's method of enquiry shows that there was not a homogeneity of Victorian thought. People read and responded to *Vestiges* differently according to their geographical, political, economic or religious circumstances. It wasn't just "Darwin's century", to use the well-known phrase of Loren Eiseley. To show all this coherently and intelligibly, as Secord has done, is no small undertaking. Take a lead from Secord's book: read *Victorian Sensation* and see what you make of it. ■

David Oldroyd is in the School of Science and Technology Studies, University of New South Wales, Sydney 2052, Australia.

The birth of scientific reading

The social structures of science were invented to cope with an explosion of printed information.

Adrian Johns

There is a persistent myth about the origin of science. It goes like this: the Scientific Revolution happened when men such as Robert Boyle and Isaac Newton freed themselves from the medieval fixation with words and instead looked directly at things. It is a simple and powerful story that has been much repeated. But it is also quite wrong. Far from emerging from a rejection of words, in fact science originated partly from a need to master as many of them as possible.

What created this need was the development of printing in the mid-fifteenth century. By 1500, more books had been printed than had ever existed before. Different books, too. For the first time since antiquity, readers could encounter Lucretius' atomism, Dioscorides' pharmacopeia, Ptolemy's geography and Archimedes' mechanics. At the same time, the New World furnished a new stock of proclaimed facts — from the medicinal powers of tobacco to the alleged origin of syphilis — that threatened to trump those available in Aristotle. All of these claims to discovery and rediscovery appeared at the hands of a burgeoning crowd of self-appointed authors, often independent of universities and royal courts. As a result, readers faced a problem not just of quantity but of quality — one uncannily similar to that confronting today's users of the Internet. What, among the outpouring of claims and counter-claims, deserved to be accepted as truth? What should one read, and how should one read it?

Writ large, such anxiety threatened to throw the very future of civilization into doubt. French scholar Adrien Baillet warned in 1685 that "the multitude of books which grows every day" would cast Europe into "a state as barbarous as that of the centuries that followed the fall of the Roman Empire". In response, contemporaries rushed to create tools to help the beleaguered reader. The Renaissance thus saw a proliferation of typographical, lexicographical and bibliographical schemes to alleviate the massive oversupply of printed ideas. Konrad von Gesner pioneered them with his mammoth *Universal Library* (1545). Gesner sought to emulate in this multivolume work the ancient Library of Alexandria, which had aimed to garner all the books of the world. By the time his last volume appeared, Gesner's

'library' contained descriptions of some 15,000 works but was still incomplete. Others suggested note-taking methods, routines for memorizing, and procedures for recalling what had been memorized. Still others mooted technological solutions, the best-known being Agostino Ramelli's book wheel. When rotated, this device brought a succession of open books into view, producing a kind of rude mechanical hypertext.

All these solutions had one thing in common. They presupposed that the multiplicity of printed opinions must be reduced to something apprehensible by just one pair of eyes. This soon proved an impossible dream. But there was an alternative: instead of reducing all books to one, why not multiply the capacities of the reader? Sure enough, throughout the sixteenth and seventeenth centuries groups of scholars and gentlefolk began to make reading itself a collective enterprise. Casually at first, but then with increasing formality, they shared out the labour of evaluating new and reissued works among small social groups. The readings that resulted merged the perspectives and expertise of the collaborators into one. Simple as it sounds, this division of readerly labour was also rather new. It would have its most remarkable effect in the period's great learned academies — and in particular in the fountainhead of experimental science itself, the Royal Society of London.

The society took the idea of collective reading and made it a settled part of its programme. It established committees, for example, to read widely in subjects such as mechanics and husbandry, sorting the worthy from the worthless. From this it gained three benefits. First, the reading itself generated proposals for new experiments — which, after all, were the society's *raison d'être*. Second, because collective reading involved bringing varied perspectives to bear, it exemplified the society's claim to be disinterested and undogmatic. And third, it

By 1500, more books had been printed than had ever existed before.



The book wheel — ancestor of the Internet.

permitted the fellows to appraise new books on an informed basis. After all, only by knowing what was in the existing literature could the society collectively confirm a proclaimed new discovery as indeed something new. So countless would-be authors — among them Boyle and Newton — submitted their creations to the Society, where they were "perused" by committees. Their shared readings would then be announced before the Society at large, where they would give rise to experiments, conversations, correspondence and, at length, more reading. In this way collective reading became a self-perpetuating process, and one that fuelled the continuing vitality of early science.

The results of all these perusals were often themselves printed, adding still more to the ocean of books. Newton's *Principia* (1687) is the most famous volume to have emerged from the process. More likely, however, contributions would appear as papers in the *Philosophical Transactions*. This was the first scientific journal, and as such it served as the template for a horde of imitators. Some of these claimed to be 'universal bibliothèques' in their own right. They asserted that their limitless reviews and abridgements, produced through the shared reading of contributors, amounted to a dynamic synopsis of human knowledge itself. Thus did the old ambition of Gesner and the Alexandrians reappear, helping to shape a newly commercial — and a newly scientific — age. ■

Adrian Johns is in the Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125, USA.

The key to the past?

Richard B. Alley

Lacking an actual time machine for studying volcanoes, drifting continents, dinosaurs and other denizens of the past, Earth scientists have to read the histories written in the pages of sediments. To do this, geologists have to use the relation between modern processes and their sedimentary products as a Rosetta Stone to translate ancient sediments into the grand story of the Earth. This use of the present as the key to the past is called uniformitarianism — but just how uniform was the past?

The first uniformitarian was probably James Hutton, who observed in the late 1700s that the passing of centuries had worked only small changes to a Roman wall in England, whereas much larger changes had broken rocks into pebbles and sand. By ignoring the constraints of scriptures and considering the possibility of an old world, Hutton found that he could explain geology as the product of processes still acting, without recourse to the supernatural or the catastrophic.

Hutton's uniformitarian ideas were rescued from his poor writing and distributed widely by Charles Lyell during the mid-1800s. However, perhaps overreacting to the persistence of Biblical literalists, Lyell took uniformitarianism to an extreme, arguing not only that modern processes acted in the past to shape the Earth, but that those processes have always occurred at similar rates.

Lyell gave Charles Darwin the vast ages on which to stage evolution. Although thus advancing biology, Lyell may have delayed the study of the origins and early evolution of Earth, the environment and life itself by taking too literally Hutton's "no vestige of a beginning, no prospect of an end".

Lyell never observed the impact of a large meteorite or the evidence for an almost totally frozen 'snowball Earth'; most of his examples were drawn from the last eighth of

Earth's history. Our present-day, broader view of events and ages has moved us back towards Hutton's uniformity of 'physical law' or of principles of behaviour. Rates and even dominant processes may change, but the underlying order of physics, chemistry and biology remains.

Oddly enough, while deep-time geologists have been quietly moving away from the Lyellian uniformitarianism designed for them, the concept lives on in palaeoclimatology. The possibility of anthropogenic or natural climate change affecting societies and ecosystems drives a small army of researchers trying to understand and even predict the behaviour of the Earth system. The history of the system is used in construction and testing of models, and the more quantitative that history, the better. That history is written in sea-floor muds, tree rings, cave formations and other sedimentary archives. Literally hundreds of physical, chemical, isotope and biological palaeoclimate indicators are used.

The palaeoclimate technique is the essence of simplicity: (1) relate some characteristic(s) of sediment to some characteristic(s) of climate, by measuring how they vary together now in different places, or over the short times of instrumental records; (2) assume that this relation was the same at a place over different times in the past; (3) measure the appropriate characteristic(s) on older sediment; and (4) calculate the past climate characteristic(s). One needs good instruments, accurate dating techniques and the ability to sample the old sediments and to measure the right things. But with enough money, mass spectrometers, drill ships and brilliant graduate students, the procedure is straightforward.

But the spirit of Lyell is lurking. As R. S. Bradley wrote in his influential text: "It is assumed that the modern relationships observed have operated, unchanged, throughout the period of interest (the principle of uniformitarianism)."

Uniformitarianism

"Until we understand more of past climates, some of the most modern Earth science will continue to rely on one of the oldest geological assumptions."

Sometimes, the uniformitarian hypothesis is easy to believe. The ice of Greenland is coldest a kilometre down because it has not finished warming from the last ice age. Simple assumptions, such as that past ice and present ice have similar thermal properties, and that space aliens have not been installing chilling units, then allow us to constrain Greenland ice-age temperatures.

But in other cases the uniformitarian hypothesis is harder to justify. Today, the average precipitation in colder places is isotopically lighter, with a highly reproducible slope. This relation has often been used to estimate past temperatures at a place. But when this substitution of space for time was tested against the temperature of Greenland's ice sheet, the isotope thermometer erred by a factor of more than two. Similarly, early work applying recent spatial patterns of ocean-sediment shell types to past temperatures missed most of the significant ice-age cooling of the tropical Pacific, affecting our ability to understand the glacial world.

The people applying the uniformitarian hypothesis to climate change know this well, and are rapidly developing new indicators. Agreement of independent indicators is harder to dismiss, as when two new palaeothermometers based on trapped gases in Greenland ice agreed beautifully with the temperatures in the ice sheet. New indicators often rely on physically simple systems, avoiding the confounding effects of other influences. Until we understand more of past climates, some of the most modern Earth-science research will continue to rely heavily on one of the oldest Earth-science assumptions, that of uniformitarianism. ■

Richard B. Alley is in the Environment Institute and Department of Geosciences, Pennsylvania State University, 204A Deike Building, University Park, Pennsylvania 16802, USA.

FURTHER READING

Alley, R. B. *The Two-Mile Time Machine: Ice Cores, Abrupt Climate Change, and Our Future* (Princeton Univ. Press, 2000).

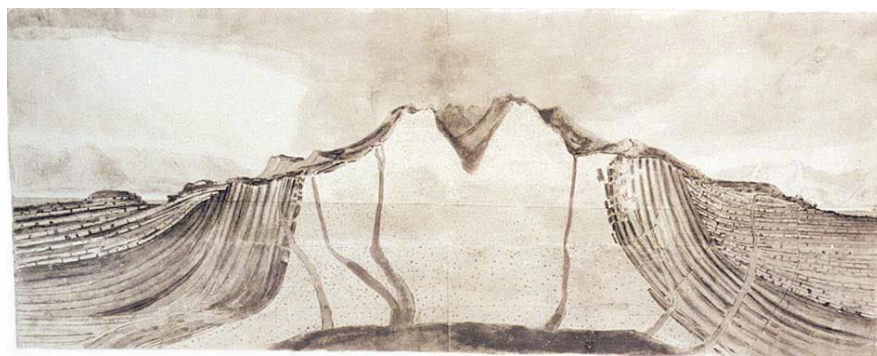
WEBLINKS

PAGES, the Past Global Changes project

► <http://www.pages.unibe.ch>

National Geophysical Data Center

► <http://www.ngdc.noaa.gov/paleo/paleo.html>



Hutton's geological observations, such as this watercolour, paved the way for uniformitarianism.

Talk is cheap in the city

Henry L. Bertoni

Telecommunications companies have paid a heavy price for their share of the radio spectrum. So they have been quick to exploit 'multiple antennas' that can increase transmission rates in urban areas.

Because we love to talk, and because modern microchips have brought down the cost, cellular (mobile) phones have become a ubiquitous feature of Western society since their introduction in the early 1980s. This demand has made the wireless-communications industry a leading part of the high-tech economy. But growth depends on attracting new subscribers, and the market is already reaching saturation in some areas—Finland leads the world in this respect, with 71 cellphones per 100 people. So, in order to keep growing, the industry must find new applications for the public to buy into.

The wireless industry sees data services, particularly e-mail and Internet links, as the solution. Unfortunately, compared with wired networks, wireless connections generally suffer from a low rate of data transmission, which results in unacceptable delays when downloading a typical web page. Three solutions to this problem have been proposed, and in this issue M. R. Andrews, P. P. Mitra and R. deCarvalho describe how they are working towards the most ambitious of the three (*Nature* **409**, 316–318; 2001).

One solution developed by a consortium of large companies is to slim down the content of data-rich web pages so they work with the low data rates of cellular networks and the small screens of palm-sized devices. This approach is embodied by the wireless application protocol (WAP), which incorporates a lightweight version of the mark-up language used to format documents on the web.

A second solution limits the length of the radio transmission to short distances, over which relatively high data rates are possible. Wireless versions of local-area networks (LANs) take this approach, but require that the wireless device is within a few tens of metres from a connection to a wired network. So public areas must be dotted with these wired access points, or the user will be restricted to a small area, thereby losing all the benefits of a portable technology.

The most ambitious approach, and the one that will benefit from the ideas of Andrews and colleagues, seeks to increase the data transmission rate of the radio channel, even for the much longer distances covered by cellphones.

Radio signals from a base station or wired access point are often scattered by objects such as buildings (Fig. 1a). This scattering provides multiple paths between the base

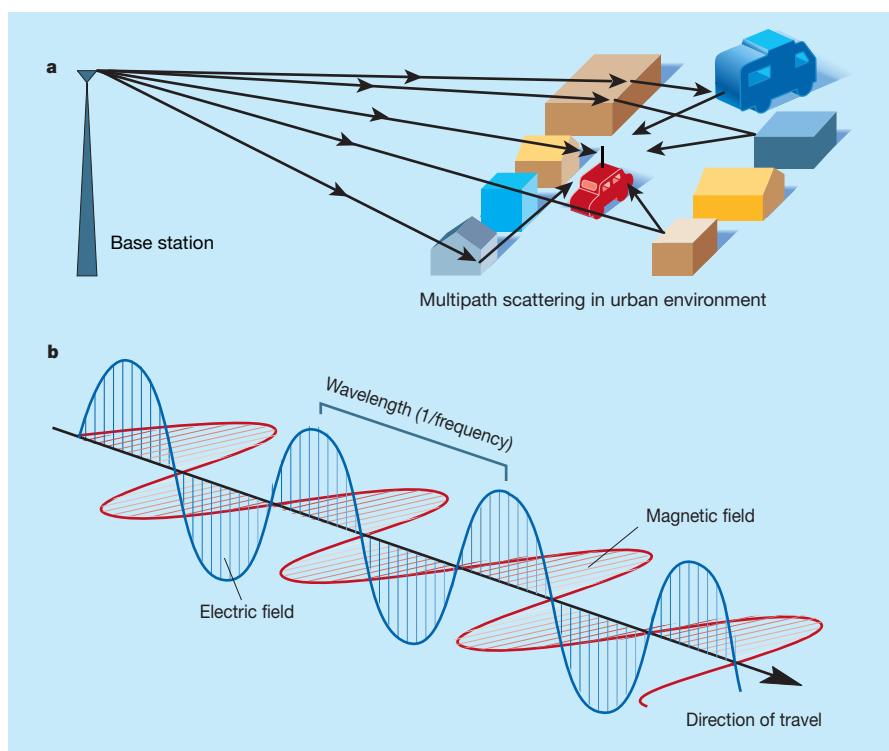


Figure 1 The smart approach to improving wireless communications. **a**, In an urban environment, the radio signals transmitted by a base station are scattered along several paths. These slightly different signals can be picked up by multiple antennas on wireless devices, resulting in greater signal reliability. **b**, Radio waves are just like any other electromagnetic wave: in free space, the electric and magnetic fields are perpendicular to each other and perpendicular to the direction of propagation — they are said to be linearly polarized. On page 316 of this issue, Andrews *et al.* show that the different polarization states created by scattering radio waves can be detected by a type of multiple antenna that is more compact than existing systems. This feature promises greater wireless capacity for small mobile devices in urban areas.

station and the wireless device, and allows them to communicate even when they are not in sight of each other. When cellular networks were first introduced, this multipath scattering gave rise to destructive interference, echoing and other problems that limited signal quality. Various approaches were taken to overcome these problems, one of which is to use multiple antennas to take advantage of the multiple communication channels created by scattering.

In view of the difficulties caused by multipath scattering for ordinary cellular networks, it came as a surprise when communications engineers suggested that multiple antennas could be used to increase the rate of data transmission in a scattering environment, such as a dense urban area. By using a group of N antennas for transmission and N for

reception, and with a signal-processing system in the receiver, they were able to transmit data at N times the rate that could be achieved using single-transmit and single-receive antennas. This improvement is made possible by the unique 'echo signature' of each of the transmitting antennas, which is detected by the signal-processing system. Such gains are not possible when the antennas are in a non-scattering environment, because N independent signals have to be transmitted in order to take advantage of the increased capacity.

The type of multiplicative scaling offered by multiple antennas has already had a big impact on the wireless industry, in which much smaller gains can generate excitement. But there is one disadvantage with this approach. As originally conceived, the multiple antennas within a group have to be

separated by a small distance, which could be as small as half the radio wavelength if they are located at the receiver, or several wavelengths if they are at the base station. Depending on the frequency band used by the wireless network, the wavelength is about one-third or one-sixth of a metre. As a result, the use of multiple antennas is restricted to base stations, because they cannot be placed far enough apart in palm-sized devices.

This is where Andrews and colleagues come in. They show that it is not necessary for the antennas to be physically separated to obtain an increase in the rate of data transmission. Instead, the antennas can be located at the same point as long as they receive signals with different polarization states. In an electromagnetic wave, such as a light beam or radio signal, the vibrating electric and magnetic fields that make up the wave can point in different directions, known as polarization states. In free space, there are only two polarization states because the electric and magnetic field must be perpendicular to one another and to the wave's direction of propagation (Fig. 1b).

In a scattering environment, however, extra polarization states are created in all three directions in space at the receiver. Andrews and colleagues exploit this fact to triple the data rate in a simple system by using a group of three perpendicular electric-dipole antennas, all located at the same point. Each electric-dipole antenna transmits or receives a component of the electric-field polarization in three dimensions. What is more, the authors show from theoretical simulations that they could increase the data rate sixfold ($N=6$) by using three magnetic-dipole antennas at the same location. For this to work, the three magnetic-field components must be uncorrelated with each other and with the three components of the electric field, in order for there to be six independent information channels. This is a surprising finding because in free space the magnetic field can always be predicted if the electric field is known.

In their elegant experiment, Andrews and colleagues have unequivocally demonstrated that electric-dipole antennas increase data rates by accessing different polarization states. But when developing mobile devices it is likely that designers will use other types of antennas, such as patch antennas, that can more easily be integrated into the case of the terminal. Nonetheless, the knowledge that multiple antennas located at the same place can increase data capacity sixfold will encourage engineers to be more aggressive when developing integrated antennas for wireless services requiring high rates of data transfer.

Henry L. Bertoni is in the Department of Electrical and Computer Engineering, Polytechnic University, Brooklyn, New York 11201, USA.
e-mail: hbertoni@poly.edu

Diabetes

The missing link with obesity?

Jeffrey S. Flier

The molecular mechanisms underlying the link between obesity and diabetes have been elusive. A new protein, christened 'resistin', can now be added to the panoply of factors that may be involved.

Type II diabetes is the most common form of diabetes in the Western world, and is strongly linked to obesity — over 80% of sufferers are obese. The molecular basis for this link has remained a mystery. On page 307 of this issue¹, however, Stepan and colleagues describe how they have identified a new hormone, which they have named 'resistin', that is produced by fat cells. Their results indicate that resistin may form at least part of the missing link between obesity and diabetes.

In patients with type II diabetes, insulin is less able to promote the uptake of glucose into muscle and fat, and to inhibit the production of glucose by the liver. Several molecular mechanisms are known by which this state — insulin resistance — comes about, but the full picture is not yet available².

Obesity is characterized by the increased storage of triglycerides (fat molecules) in adipose tissue and causes insulin resistance. But how does this increased energy storage in fat cells (adipocytes) promote insulin resistance in muscles, the liver and else-

where in the body? For many years, it looked as if free fatty acids would provide the link. These products of triglyceride metabolism are the main form in which energy is transferred from stores in adipose tissue to other sites in the body for metabolic use. After all, levels of free fatty acids in the bloodstream are higher in obese than in non-obese people, and free fatty acids can induce insulin resistance in tissues other than adipose tissue³. However, as well as having a role in energy storage, adipocytes also secrete numerous peptides that might lead to insulin resistance or other complications of obesity (Fig. 1).

Two such peptides are the cytokine tumour-necrosis factor- α (TNF- α) and the hormone leptin. Although better known for its roles in inflammation and immunity, TNF- α is expressed in normal adipocytes, is overexpressed in adipocytes from obese people, and can cause insulin resistance through effects on insulin-mediated cellular signalling pathways⁴. TNF- α is almost certainly involved in insulin resistance *in vivo*.

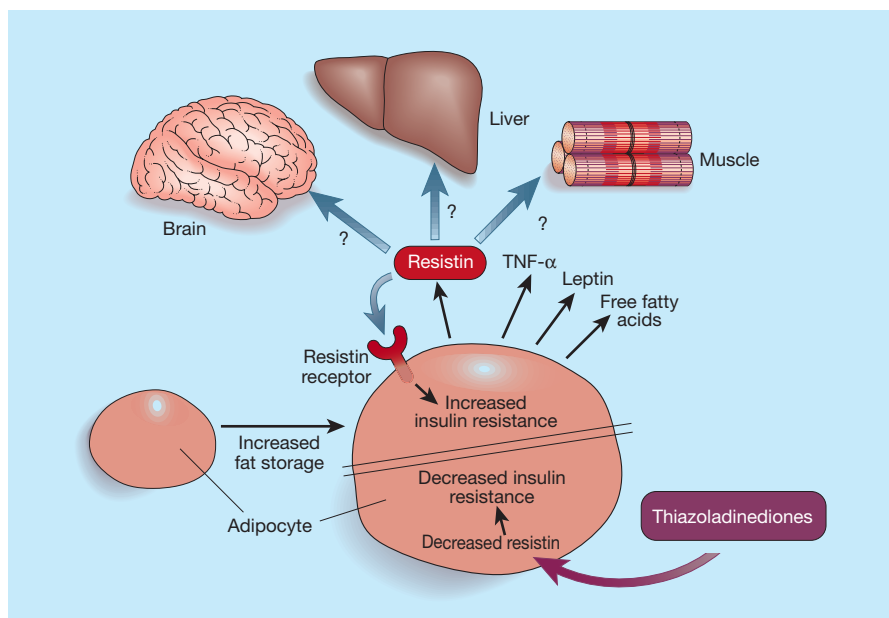


Figure 1 Obesity causes certain tissues in the body (such as muscle and liver) to be less sensitive to insulin. Such insulin resistance is one of the main features of type II diabetes. From left, as fat cells (adipocytes) store more fat molecules and enlarge, they release several products that can modify the body's sensitivity to insulin. Free fatty acids and tumour necrosis factor- α (TNF- α) cause insulin resistance, and leptin, which regulates energy balance, probably causes insulin sensitivity. Stepan *et al.*¹ have identified a new protein, resistin, that is secreted by adipocytes. Resistin causes insulin resistance through its effects on adipocytes and perhaps other tissues. Thiazolidinedione drugs reduce insulin resistance and are used to treat type II diabetes. These drugs suppress the expression of resistin by adipocytes, and their antidiabetes effects may, at least in part, be achieved through this mechanism.

But the balance of evidence suggests that other factors must be required, too.

Is leptin such a factor? It is best known as the adipocyte hormone whose absence causes profound obesity in rodents and humans, but it also has potent effects on insulin's action^{5,6}. In rodents, for example, an inherited deficiency in leptin causes both severe insulin resistance and obesity. Returning leptin to these rodents reverses insulin resistance, but by mechanisms that are independent of effects on food intake and body weight⁷. This reversal in insulin resistance may result from leptin working within the brain as well as directly on target tissues. But in contrast to rodents, in which leptin deficiency causes obesity, levels of leptin are higher than normal in most obese people⁸. So it is not yet clear how alterations in leptin might affect insulin resistance.

Steppan *et al.*¹ have now taken a creative approach to identifying new adipocyte-derived mediators of insulin resistance. The thiazoladinedione class of antidiabetes drugs reduces insulin resistance by acting through a nuclear receptor protein that is abundant in fat cells⁹. This protein is the peroxisome proliferator-activated receptor- γ (PPAR- γ)¹⁰, which has been found to guide the differentiation of adipocytes¹¹. The results of genetic and pharmacological experiments revealed that this nuclear receptor also affects insulin sensitivity, by unknown mechanisms that presumably involve altered gene expression in adipocytes. Might the thiazoladinediones work through PPAR- γ to switch on or off an adipocyte-specific gene that is involved in insulin-mediated signalling pathways? Steppan *et al.* assumed that it does.

By treating a differentiated adipocyte cell line with thiazoladinediones, the authors have identified a new messenger RNA that is expressed only in adipose tissue and is suppressed by thiazoladinediones. The authors find that the protein encoded by the new adipocyte-specific mRNA is overexpressed in fat in a variety of rodent models of obesity. Moreover, thiazoladinediones reduce the amount of this protein that is secreted from adipocytes *in vitro* and released into the bloodstream in mice.

The evidence that this new protein is involved in insulin resistance is provocative, but, so far, circumstantial. So, to test this hypothesis directly, the authors administered recombinant protein, and antibodies that recognize the endogenous protein, to fat cells *in vitro* and to mice. Some of the effects were rather small, but they all point in the same direction — that the protein, named resistin, is an important link between obesity and diabetes, and may mediate the beneficial effects of a major class of antidiabetes drugs.

Nonetheless, questions remain. Steppan *et al.* have shown that resistin suppresses

insulin's ability to stimulate glucose uptake into adipose cells. But it is not known whether resistin also acts on important physiological targets, such as muscles, the liver or even the brain. Resistin is presumably detected in the bloodstream by a receptor protein on the surface of target cells. But what is the identity of the resistin receptor? How exactly does resistin-mediated signalling antagonize insulin-mediated signalling and insulin's effects? And what is the role of resistin in normal physiology? Answers to these questions will probably require gene-knockout experiments, careful examination of physiological responses to administration of resistin, and identification of the factors that naturally regulate the amount of resistin in the blood. It will also be important to determine whether genetic differences in people's ability to produce or respond to resistin contribute to variations in insulin sensitivity and susceptibility to diabetes.

More generally, however, this work¹ illus-

trates an important principle of modern molecular science. Drugs that are discovered empirically, such as thiazoladinediones, can often be used for much more than treating diseases. In the hands of good scientists, they can also be essential tools for discovering previously hidden molecular pathways. ■

Jeffrey S. Flier is at the Beth Israel Deaconess Medical Center, Research North 325, 99 Brookline Avenue, Boston, Massachusetts 02215, USA.
e-mail: jflier@caregroup.harvard.edu

1. Steppan, C. M. *et al.* *Nature* **409**, 307–312 (2001).
2. Kahn, B. B. & Flier, J. S. *J. Clin. Invest.* **106**, 473–481 (2000).
3. Boden, G. *Proc. Assoc. Am. Physicians* **111**, 241–248 (1999).
4. Hotamisligil, G. S., Peraldi, P. & Spiegelman, B. M. *Curr. Opin. Endocrinol. Diab.* **3**, 16–23 (1996).
5. Friedman, J. M. & Halaas, J. L. *Nature* **395**, 763–770 (1998).
6. Ahima, R. S. & Flier, J. S. *Annu. Rev. Physiol.* **62**, 413–437 (2000).
7. Shimomura, I., Hammer, R. E., Ikemoto, S., Brown, M. S. & Goldstein, J. L. *Nature* **401**, 73–76 (1999).
8. Considine, R. V. *et al.* *New Engl. J. Med.* **334**, 292–295 (1995).
9. Olefsky, J. M. *J. Clin. Invest.* **106**, 467–472 (2000).
10. Lehmann, J. M. *et al.* *J. Biol. Chem.* **270**, 12953–12956 (1995).
11. Tontonoz, P. *et al.* *Genes Dev.* **8**, 1224–1234 (1994).

Geochemistry

New prospects for old gas

Bernard Marty

Isotope studies furnish evidence of the source of CO₂ in certain natural-gas reserves, and of the long-term retention of such gas in unexpected environments such as ancient continental crust.

How long can natural gases be retained in geological formations deep underground? From a study of carbon and helium in such gases, described by Ballentine *et al.* on page 327 of this issue¹, it seems that they can be stored for much longer than has generally been assumed. This finding has implications for identifying and exploiting underground fluids, notably new reserves of hydrocarbons such as methane.

Hydrocarbons are generated in specific geological environments, the source rock formations. They then migrate underground until they encounter a natural barrier, the cap rocks, which prevent further movement upwards. There they may remain until the barrier's geometry is changed by tectonic activity and they are released. In a tectonically stable area, the limiting factor for hydrocarbon storage is believed to be their rate of diffusion through the cap rocks. In geological terms this diffusion is thought to be swift, so the gas should not be retained for long geological periods. But Ballentine *et al.* argue that in the area they studied, hydrocarbons, along with CO₂ and rare gases of different origins, have been accumulating and stored for up to 300 million years. The region in question consists of limestone formations known as the west Texas Permian basin. The researchers conclude that continental areas such as this, which are old and

stable, could be fruitful targets for hydrocarbon exploration.

Accumulations of hydrocarbons, especially natural-gas reservoirs, often contain CO₂. This can lower the economic value in two ways. If there is a lot of CO₂, there will be less hydrocarbon, and there are additional costs involved in separating the two. So the aim of oil companies is to locate high-quality deposits of natural gas in the Earth's crust, for instance by identifying the origin of gaseous components and how they were incorporated into the hydrocarbon reservoirs. There are two possible origins for the CO₂: it may have been derived from the underlying mantle through degassing of magmas; or it may stem from thermal degradation of carbon-rich rocks, such as limestones, in the crust.

To find out which is the case, analysis of the isotopic composition of the carbon present is not very helpful. Carbon has only two isotopes, and the isotopic signature of mantle carbon lies between those of carbonate (limestone) and of organic carbon (methane, for example). Likewise, the chemical composition of natural gases is of limited value. It allows the conditions in which the gases were generated to be inferred, but not the origin of their component elements. Fortunately, natural-gas deposits contain rare gases, such as helium,

which have proven to be good fingerprints of origin. They are inert, and so are not involved in the chemical speciation that occurs in hydrocarbon formation. And their source — whether in the atmosphere, ground water, continental crust or mantle — can be identified reasonably well from their isotopic composition.

So, together with the relative proportions of He to CO₂, helium isotopes provide a unique way to identify the origin of CO₂ (refs 2, 3). The isotopic composition of helium, ³He/⁴He, in the mantle is three or four orders of magnitude higher than that of pure crustal gases. This is because primordial ³He trapped during the Earth's formation is still present in the mantle, whereas the crust is degassed in mantle volatiles and is rich in radiogenic ⁴He produced by the radioactive decay of uranium and thorium isotopes. As a result, the CO₂/³He ratio of the mantle is (1–2) × 10⁹ whereas that of crustal gases in stable continental areas is much higher — around 10¹¹–10¹³.

Natural gases from the west Texas Permian basin contain from a few per cent to as much as 97% CO₂. Ballentine *et al.*¹ have made precise measurements of the ³He/⁴He and CO₂/³He ratios in gases from several wells in the area, in an attempt to determine the origin of the CO₂ and explore the paths and processes by which it became mixed with the hydrocarbons. They found ³He/⁴He ratios higher than those characteristic of radiogenic production in the crust, and CO₂/³He ratios close to the mantle ratio, much lower than typical crustal ratios. So they conclude that this CO₂ is magmatic in origin.

But how are gases originating in the Earth's mantle transferred into the crust? The most plausible process is magmatism — roughly speaking, deep-seated volcanic activity — and Ballentine *et al.* have attempted to identify the potential areas of magmatism that have contributed CO₂ to the west Texas Permian basin. Magmatism occurred during the development of the Basin and Range Province, which lies 100 km or more west of the Texas basin. But Ballentine *et al.* argue that this is an unlikely source because the CO₂/³He mantle signal increases towards the south of the Texas Permian basin, suggesting a different pathway for gas migration. Instead, they propose that mantle-derived CO₂ entered the basin during extension and uplift of the area around 300 million years ago, which is believed to be the last large tectonic episode to affect the region. In their view, hydrocarbon generation began 280 million years ago — that is, after CO₂ had entered the basin.

Several reports indicate that natural gases, especially hydrocarbon-rich gases, are often associated with volatiles derived from the mantle. But such an association generally occurs in basins that have experienced tec-

tonic activity comparatively recently (during the Cenozoic, the era running from 65 million years ago until the present). Examples of such basins are those that have been affected by Alpine tectonics, such as the Rhine graben or the Pannonian basin in Hungary^{3,4}, or those formed in subduction zones where tectonic plates slide down into the mantle, such as the Green Tuffs in Japan^{5,6}. If Ballentine and colleagues' view that gases can be stored in much older, currently stable continental areas is correct, it means that there may well be appreciable amounts of hydrocarbons in such regions waiting to be discovered and exploited.

More generally, their conclusion supports the geochemical view that fluids (ground water, as well as natural gas) can be stored in geological basins for very long periods of time, rather than the results of hydrodynamical modelling which suggest that underground fluids move compara-

tively rapidly in geological terms. The key issues in these debates are the variation in permeability of geological barriers and the diffusivity of fluid species through such barriers — both parameters are only poorly known and are difficult to assess over geological time periods. Further case studies that address these matters would be most welcome.

Bernard Marty is in the Centre de Recherches Pétrographiques et Géochimiques, 15 Rue Notre-Dame des Pauvres, BP 20, 54501 Vandoeuvre les Nancy Cedex, France.

e-mail: bmarty@crpg.cnrs-nancy.fr

1. Ballentine, C. J., Schoell, M., Coleman, D. & Cain, B. A. *Nature* **409**, 327–331 (2001).
2. Marty, B. & Jambon, A. *Earth Planet. Sci. Lett.* **83**, 16–26 (1987).
3. O'Nions, R. K. & Oxburgh, R. O. *Earth Planet. Sci. Lett.* **90**, 331–347 (1988).
4. Mamyrin, B. A. & Tolstikhin, I. N. *Helium Isotopes in Nature* (Elsevier, Amsterdam, 1984).
5. Poreda, R. & Craig, H. *Nature* **338**, 473–478 (1989).
6. Sano, Y. & Wakita, H. *J. Geophys. Res.* **90**, 8729–8741 (1985).

RNA silencing

Diced defence

David Baulcombe

RNA silencing allows cells to block invading viruses or mobile DNAs. An RNA-cleaving enzyme involved in the first step of silencing has now been identified.

Several years ago, I heard a famous geneticist describe the difference between talks given by geneticists and those by biochemists. To him, a geneticist's talk was one with ideas but no data; a biochemist's talk was just the opposite. I doubt that such an outrageous suggestion would be made nowadays, because so many ideas and data involve a combination of genetics and biochemistry. One such example is provided on page 363 of this issue¹, where Bernstein and colleagues describe how they have identified — in the fruitfly *Drosophila melanogaster* — an enzyme that they call 'Dicer' because it chops RNA into small pieces of uniform size.

Bernstein *et al.* were investigating a process known as RNA interference or RNA silencing. This phenomenon provides organisms with a defence against mobile DNA elements, which cause mutations when they insert themselves within, or close to, a gene². In plants, RNA silencing may also offer protection against viruses³. The existence of RNA silencing, at least in animals, has been inferred from experiments in which double-stranded (ds)RNA was introduced into cells⁴. Most RNA in a cell is single-stranded, so the presence of dsRNA might signal to the cell that it is being invaded by mobile DNA or viruses.

RNA is cut up into smaller chunks during at least two steps of RNA silencing. First,

dsRNA is processed into many short pieces, each of about 22 nucleotides⁵. These fragments are then incorporated into a multi-subunit complex, which carries out the second RNA-degradation reaction. Here, the target RNA is single-stranded messenger RNA⁶. Apparently, the role of the 22-nucleotide RNAs generated in the first step is to interact, by base pairing, with mRNA in which the sequence of nucleotides is the same as in the dsRNA. The short RNA molecules thus guide the RNA-degrading enzyme (the RNase) part of the multi-subunit complex, ensuring that it degrades only those mRNA species that are related to the dsRNA.

An advantage of this two-stage process is that each molecule of dsRNA primes several RNase molecules with a guide RNA. Consequently the cell can mount a large response to only a few dsRNA molecules. This feature is ideal for defence against viruses or mobile DNAs that produce both single- and double-stranded RNA. The virus would be arrested early in an infection and the mobile DNA would be frozen before it had damaged the genome.

Bernstein and colleagues¹ now show that different RNases are required for the two RNA-degrading steps. They give the name Dicer to the enzyme needed for the first step. The authors started their search for this enzyme by assuming that it would be similar

to the known RNases that are specific for dsRNA. These enzymes are mostly members of the RNase III family, and have characteristic motifs in their protein sequence. Bernstein *et al.* identified candidate enzymes by scanning the *Drosophila* genome for genes encoding proteins with these RNase III signatures.

The authors then used three tests to confirm that one of these genes, *dicer*, encodes the RNA-processing enzyme needed for the first step of RNA silencing. First, they showed that *dicer* encodes a protein that processes dsRNA into 22-nucleotide fragments *in vitro*. Second, antibodies raised against this protein recognized an enzyme in *Drosophila* extracts that process dsRNA into 22-nucleotide fragments. Third, when Bernstein *et al.* introduced *dicer* dsRNA into *Drosophila* cells, they found that the cells showed a diminished ability to carry out RNA silencing of a gene encoding green fluorescent protein. In effect, the authors were using RNA interference to suppress a gene required for RNA interference. It would not be possible to use RNA interfer-

ence to suppress *dicer* completely if this gene is indeed required for silencing. But, as observed by Bernstein *et al.*, it was possible to get partial suppression.

One feature of the amino-acid sequence of Dicer — in addition to the RNase III motifs — is a sequence known as the PAZ motif⁷. This sequence is also present in the *Drosophila* ARGONAUTE protein and related proteins in the thale cress *Arabidopsis thaliana*⁸, the nematode worm *Caenorhabditis elegans*⁹ and the fungus *Neurospora crassa*¹⁰. These proteins do not have the RNase III motifs, but they are nevertheless required for RNA silencing. A clue to the role of the PAZ motif comes from Bernstein *et al.*'s unpublished data. Apparently, the multi-subunit RNA-processing complex needed for the second RNA-silencing step in *Drosophila* contains a relative of ARGONAUTE; perhaps Dicer and this protein interact through their PAZ domains (Fig. 1). The interaction would explain how the 22-nucleotide RNAs produced by Dicer could be incorporated into the multi-subunit complex that degrades single-stranded RNA.

No mutants of *dicer* have yet been identified. But mutation of the *CARPEL FACTORY* gene — a relative of *dicer* in *Arabidopsis* — leads to plants that grow abnormally and have defective flowers¹¹. Strikingly, this mutation and others that affect both RNA silencing and growth^{8,12} all affect either animal germline cells (from which reproductive cells form) or their plant equivalent, meristems. A defensive RNA-silencing system would be especially important in these cells,

because a virus infecting such a cell would be transmitted from one generation to the next. Similarly, mutations induced in these cells by mobile DNA would result in heritable genetic instability.

But why should loss of a defensive RNA-silencing system result in defective growth of these mutants? One possibility is that RNA silencing is also involved in regulating genes required in the germ line for normal growth and development. Alternatively, the growth defects might be a side effect of the loss of the defence mechanism in germline cells. A combination of biochemical and genetic approaches, laced with ingenuity, will be needed to dig further into the mechanism and role of RNA silencing in defence and growth.

David Baulcombe is in the Sainsbury Laboratory, John Innes Centre, Norwich NR4 7UH, UK.
e-mail: david.baulcombe@bbsrc.ac.uk

1. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. *Nature* **409**, 363–366 (2001).
2. Plasterk, R. H. A. & Ketting, R. F. *Curr. Opin. Genet. Dev.* **10**, 562–567 (2000).
3. Mourrain, P. *et al.* *Cell* **101**, 533–542 (2000).
4. Fire, A. *et al.* *Nature* **391**, 806–811 (1998).
5. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. *Cell* **101**, 25–33 (2000).
6. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. *Nature* **404**, 293–296 (2000).
7. Cerutti, L., Mian, N. & Bateman, A. *Trends Biochem. Sci.* **25**, 481–482 (2000).
8. Fagard, M., Boutet, S., Morel, J.-B., Bellini, C. & Vaucheret, H. *Proc. Natl Acad. Sci. USA* **97**, 11650–11654 (2000).
9. Tabara, H. *et al.* *Cell* **99**, 123–132 (1999).
10. Catalanotto, C., Azzalin, G., Macino, G. & Cogoni, C. *Nature* **404**, 245 (2000).
11. Jacobsen, S. E., Running, M. P. & Meyerowitz, E. M. *Development* **126**, 5231–5243 (1999).
12. Smardon, A. *et al.* *Curr. Biol.* **10**, 169–178 (2000).

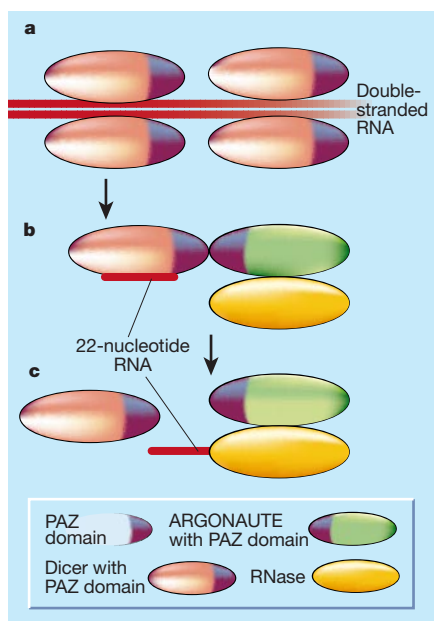


Figure 1 According to one model for RNA silencing, there are two stages during which RNA is cleaved. **a**, In the first stage in *Drosophila*, the enzyme Dicer (identified by Bernstein *et al.*¹) binds to double-stranded RNA produced by a virus or by mobile DNA, or introduced experimentally. Dicer cleaves the double-stranded RNA into fragments of 22 nucleotides each. **b**, Dicer then associates through its so-called PAZ domain with a relative of ARGONAUTE, another PAZ-domain-containing protein. **c**, This association would allow the 22-nucleotide RNA to be transferred to an RNase associated with the ARGONAUTE relative, guiding the RNase to single-stranded messenger RNAs that match the 22-nucleotide RNA. The RNase would then cleave the single-stranded RNAs (not shown).

Neutron stars

Pulsars crash the magnetar party

Jim Cordes

The discovery of a youthful neutron star with an extreme magnetic field blurs the once clear divide between magnetars and pulsars.

Radio pulsars are rapidly spinning, strongly magnetized neutron stars that produce pulses of X-rays and γ -rays as well as radio waves. A combination of data from two X-ray satellites recently led to the discovery of a very young pulsar with an unusually high magnetic field, as reported by Eric Gotthelf and collaborators¹ in *Astrophysical Journal Letters*. This X-ray pulsar is associated with the supernova remnant Kesteven 75. Although neutron stars are thought to be created in supernova explosions, there are surprisingly few clear examples of this. Gotthelf *et al.* suggest that the unusual X-ray pulsar may be a missing link between different classes of pulsars, and that it could provide important clues to how particles are created and radiate in the magneto-

spheres surrounding neutron stars. It may turn out that many of the neutron stars in our Galaxy are born with properties similar to this pulsar rather than to the bulk of previously known neutron stars.

Neutron stars are extremely dense objects: they have masses similar to the Sun but diameters of only about 20 km. Most neutron stars are detected as 'ordinary' radio pulsars with magnetic fields of 10^{12} gauss — one trillion times as strong as the Earth's magnetic field — and spin periods between 16 milliseconds and 8.5 seconds. The rapid rotation combines with the strong magnetic field to produce electric forces, like the alternator in a car, that generate particles moving at the speed of light. These relativistic particles radiate intense electromagnetic

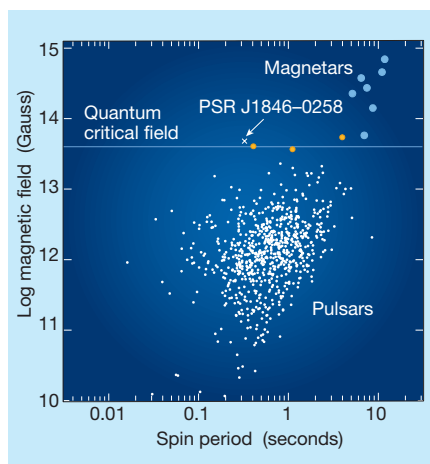


Figure 1 The distribution of magnetic field strengths and spin periods for ordinary radio pulsars (white dots) and magnetars (blue circles). Magnetars, which comprise both the soft- γ -ray repeater sources and the anomalous X-ray pulsars, have magnetic field strengths above the so-called quantum critical field. The new X-ray pulsar found by Gotthelf *et al.*¹ (PSR J1846-0258) has a magnetic field comparable to three recently discovered radio pulsars⁷ (orange circles). PSR J1846-0258 is not known to be emitting radio waves, although this may mean that the rotating radio beam is simply not directed towards Earth. The lack of known objects in the upper left corner is partly due to the fact that objects born there rapidly lose rotational energy and move to the right as they age. But it is conceivable that many neutron stars start their lives there.

waves directed along the magnetic poles, which appear to an observer as pulses of radiation as the star rotates, like the light from a lighthouse. The magnetic field also slows the pulsar down through magnetic braking, an effect caused by the radiation carrying away the star's angular momentum. Ordinary pulsars stay radio 'loud' for about 10 million years, the time it takes for them to slow down to a spin rate at which particle creation stops.

Over the past five years, considerable interest has focused on objects that appear to be even more highly magnetized than typical radio pulsars. These are the so-called magnetars with magnetic fields that range from a few times 10^{13} to nearly 10^{15} gauss. A few have been identified in X-ray and γ -ray observations (Fig. 1). They spin down much more rapidly than radio pulsars, on timescales of about 10,000 years. Two subclasses of magnetars have been identified: soft- γ -ray repeaters (SGRs) and anomalous X-ray pulsars (AXPs).

The first three SGRs were discovered in 1979: two in the plane of our Galaxy, the third as a spectacular burst of γ -rays from an object associated with a supernova remnant in a satellite galaxy of the Milky Way. The γ -ray burst was followed by a fainter oscillation indicative of a slowly spinning object,

now interpreted as a rotating neutron star that has undergone severe magnetic braking. Since 1979, one more SGR has been found, also in the plane of the Milky Way².

The AXPs are anomalous in that their spin properties differ from the much larger class of X-ray pulsars found in binary systems. The AXPs also differ from the SGRs in that they do not show the same 'bursting' behaviour. Binary X-ray pulsars tend to spin faster as they gravitationally attract material from their companion star. At present, none of the magnetars is known to be in a binary system, and their evolution is towards longer spin periods — more like that of radio pulsars. Unlike radio pulsars, however, AXPs appear to radiate more X-ray energy than they need to lose as they slow down. It is thought that some of this anomalous X-ray emission comes from decay of the strong magnetic field.

When first identified, magnetars appeared to be a minority class of neutron star, with properties distinct from those of most radio pulsars^{3,4}. But the short lifetimes of magnetars means they are probably under-represented in observed samples, suggesting that their descendants litter our Galaxy. Nonetheless, there is no consensus about the conditions that determine whether a supernova explosion will leave behind an ordinary radio pulsar or a magnetar. One possibility is that the spin rate of the neutron star at birth governs the strength of additional magnetic fields generated through any dynamo effects⁵.

It has also been suggested that magnetars differ from radio pulsars, as far as processes in the magnetosphere are concerned. In a regular pulsar, some of the relativistic particles are electron-positron pairs generated from high-energy photons. The pairs generate, in turn, more high-energy photons, thus creating a cascade. The cascade subsides when a pulsar's spin slows down. But these cascades may also be suppressed in magnetic fields larger than the so-called critical quantum field of 4.4×10^{13} gauss (ref. 6). In such intense fields a competing process — photon splitting — can degrade the photon energies to below the requirement for pair production. For a short while, the magnetars and known radio pulsars were thought to fall on opposite sides of this divide.

X-ray emission from the new pulsar found by Gotthelf *et al.*, PSR J1846-0258, is not unlike that seen from other radio pulsars, suggesting that it originates from electron-positron pairs, despite having a magnetic field greater than the critical quantum field. There are now several radio pulsars that are close to this limit⁷ (Fig. 1). Such objects seem to indicate that photon-splitting is not a dominant process in the magnetospheres of neutron stars. But our calculations of magnetic field strengths make several assumptions that may not hold for all neutron stars,



100 YEARS AGO

Mr. Hovenden has set himself the modest task of overthrowing, in the space of about 300 pages, all existing physical tenets, and substituting in their place a remarkable theory of his own. In this effort he has not succeeded, except, apparently, to his own complete satisfaction. In the first part of the book the author quotes freely from Maxwell and others, and endeavours to prove that their reasoning is fallacious. His arguments only show that he does not understand what he quotes, and that he has not appreciated the most elementary principles of the subject, such, for example, as the difference between mass and weight. Having, as he considers, sufficiently disposed of the views held by modern men of science, Mr. Hovenden proceeds to the elucidation of his own theory. It is impossible to regard this part of the book seriously, Mr. Hovenden's deductions from experiments being altogether too extravagantly absurd. It is interesting to note that his treatment of the subject is throughout entirely qualitative; we venture to think that in no single instance would Mr. Hovenden's explanations stand the test of quantitative examination.

From *Nature* 17 January 1901.

50 YEARS AGO

In a recent review of Prof. Soddy's book, "The Story of Atomic Energy" (London, 1949), Prof. Paneth directs attention to a part of the book in which "credit is given not to John Dalton but to William Higgins for being the first to proclaim the modern atomic theory, a statement supported by detailed references to two books of this 'almost forgotten investigator'". Prof. Paneth adds that "Dalton's astonishing limitations as a scientific thinker — as revealed in his notes about the constitution of gases and his stubborn fight against Gay-Lussac's results which was the consequence — seem to justify the more modest position allotted to him here. In any event, writers of chemical text-books will now have to re-examine this question." Prof. Soddy had "hoped that future historians of the atomic theory may give more attention to these two books of William Higgins. They certainly serve to establish the writer's view that, once Lavoisier's Theory of Combustion was accepted, the next natural step was the modern atomic theory, and that Higgins definitely took this step 15 years before Dalton."

From *Nature* 20 January 1951.

so there may be systematic errors in the estimated field strengths. Despite this caveat, discoveries such as PSR J1846-0258 blur the division between pulsars and magnetars. It seems entirely possible that they represent extremes of the neutron-star population, depending on their original spin and magnetic field strength. So, for neutron stars, there may well be unity in their diversity. ■

Jim Cordes is in the Astronomy Department, Cornell University, Ithaca, New York 14853-6801, USA.

e-mail: cordes@spacenet.tn.cornell.edu

1. Gotthelf, E. V., Vasisht, G., Boylan-Kolchin, M. & Torii, K. *Astrophys. J. Lett.* **542**, 37–40 (2000).
2. Woods, P. M. *et al. Astrophys. J. Lett.* **519**, L139–L142 (1999).
3. Duncan, R. & Thompson, C. *Astrophys. J. Lett.* **392**, L9–L13 (1992).
4. Kouveliotou, C. *et al. Nature* **393**, 235–237 (1998).
5. Wheeler, J. C., Yi, I., Hoflich, P. & Wang, L. *Astrophys. J.* **537**, 810–823 (2000).
6. Baring, M. G. & Harding, A. K. *Astrophys. J. Lett.* **507**, L55–L58 (1998).
7. Manchester, R. N. *et al. in Pulsar Astronomy — 2000 and Beyond* ASP Conf. Vol 202 (eds Kramer, M., Wex, N. & Wielebinski, R.) 49–54 (2000).

Evolution

Speciation in the round

David B. Wake

‘Ring species’ occur when one species grades into two at the overlap of a circular population distribution. Good examples are rare, but one case has now passed some rigorous tests.

In principle, ring species should constitute “the perfect demonstration of speciation”¹. According to the classic model, such rings are composed of populations of an organism that started from a single ancestral population and are now distributed geographically in a circular manner. Neighbouring populations in the ring can interbreed, except for the two terminal, overlapping populations which cannot and behave as if they are distinct species. The process is a culmination of a gradual adaptation of the organism to a gradient of habitats around the ring such that, when populations meet at the terminus, species-level divergence has occurred. An especially attractive feature of ring species is that they represent a piece of history, in that various stages in the formation of species are preserved and can be studied.

For various reasons there may be no such thing as the perfect ring species. But on page 333 of this issue, Irwin *et al.*² describe how they have revisited one of the earliest putative examples — that of the greenish warbler (*Phylloscopus trochiloides*; Fig. 1), which is distributed around the margins of the Tibetan plateau in Asia and was first described over 60 years ago³. They conclude that, although this warbler ring is not perfect, it meets most of the requirements of the classic model.

Identifying ring species is complicated. First, the geographical conditions are critical: there needs to have been divergence from a single ancestral source population around some kind of barrier. An appropriate balance must be reached between the opposing forces of selection (which causes geographically neighbouring populations to diverge in their adaptations) and continuing gene flow through interbreeding (which maintains genetic continuity). Moreover, the ring of populations should be continuous, although it is almost impossible to satisfy this criterion. Ranges of species are subject to constant



Figure 1 On song — an east Siberian greenish warbler.

change — they expand and contract over time, with parts becoming separated then reconnecting to form secondary contacts. In other instances, populations become extinct, producing gaps. In the case of the greenish warbler, there is a gap in China, which seems to have been caused by human destruction of the bird’s forest habitat.

A further complication is that the two terminal populations are usually identified first, and are often named as different species. Given this initial taxonomic bias, subsequent studies may overlook the possibility that a ring species exists. In other instances, possible examples have fallen prey to changing concepts of what a species is; several former ring species are no longer treated as a single species taxonomically, but rather as groups of closely related species (circumpolar gulls are a classic example).

One of the best-studied cases is the salamander *Ensatina eschscholtzii*, which has populations around the Central Valley of California with sharply divergent forms coexisting in southern California⁴. But whether or not this is a genuine ring species is a matter of controversy. There is one break in

the distribution and another area where there is a ‘weak link’ in the chain of differentiated but intergrading forms. Also, there is debate over the most appropriate taxonomy because of different interpretations of genetic data^{5–9}.

From Irwin and colleagues’ work², however, it seems that greenish warblers satisfy most of the stringent requirements for a ring species. One form ranges from the Baltic region to western Siberia, where it spreads south and intergrades with another form along the western and southern slopes of the Himalayas (see Fig. 1 of the paper on page 334). Intergradation with other forms occurs to the east, then northward, until an eastern Siberian form meets the western Siberian form in the vicinity of the Yenisei River. There the two behave as distinct biological species, and do not interbreed. For instance, as Irwin *et al.* show, they have more complex songs than southern populations, and no longer recognize each other as potential mates — this is sexual selection as a driving force in speciation. The birds are thought to have originated in the south and moved around the Himalayas, the eastern and western populations coming into contact in the north, but remaining distinct, following the end of the last glacial epoch. This fulfils the geographical and biological requirements of a ring species.

Irwin *et al.* also collected mitochondrial DNA evidence from the various populations. The DNA sequences form two well-supported clades, one eastern and the other western. In the region of the putative ancestral populations, the clades are intermixed. Typical phylogeographic analysis^{10,11} would interpret these clades as historical entities: that is, that they arose from eastern and western isolates that diverged when they were geographically isolated, and subsequently expanded their ranges both north and south to reconnect. In the north, songs diverged such that coexisting populations behave as different species, whereas in the south birds remained sufficiently similar in song characteristics to merge and form an interbreeding unit.

This interpretation is considered, but rejected, by Irwin *et al.*². Instead, they have conducted simulations (not published) which show that antecedents of the two major clades of DNA sequences might have arisen within the same area. The sequences are not very divergent, and Irwin *et al.* argue that sorting of the two primordial lineages of sequences took place in the region in which they originated. One lineage came to dominate in the eastern region and the other in the western, as the populations expanded around the mountain mass. This interpretation runs counter to a central tenet of phylogeography, which sees history as having been recorded in the phylogeny of sequence lineages, and so is likely to be controversial.

Finally, the paper is notable for incorporating two different explanations for species formation. First, the authors take sexual

selection to be responsible for the elaboration and divergence of songs in the greenish warbler. By contrast, they believe that the similarity in body size of the northern forms stems from natural selection acting on size. Both natural selection (especially in so-called ecological speciation)^{12,13} and sexual selection^{14,15} have been the subject of close recent interest. So the work of Irwin *et al.*² makes a multifaceted contribution to the current debate on species formation.

David B. Wake is in the Museum of Vertebrate Zoology and the Department of Integrative Biology, University of California, Berkeley, California 94720-3160, USA.

e-mail: wakelab@uclink.berkeley.edu

1. Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, MA, 1963).

2. Irwin, D. E., Bensch, S. & Price, T. D. *Nature* **409**, 333–337 (2001).
3. Ticehurst, C. B. *A Systematic Review of the Genus Phylloscopus (Willow-warblers or Leaf-warblers)* (Trustees of the British Museum, London, 1938).
4. Stebbins, R. C. *Univ. California Publ. Zool.* **48**, 377–526 (1949).
5. Moritz, C., Schneider, C. J. & Wake, D. B. *Syst. Biol.* **41**, 273–291 (1992).
6. Jackman, T. R. & Wake, D. B. *Evolution* **48**, 876–897 (1994).
7. Wake, D. B. *Proc. Natl Acad. Sci. USA* **94**, 7761–7767 (1997).
8. Highton, R. *Herpetologica* **54**, 254–278 (1998).
9. Wake, D. B. & Schneider, C. J. *Herpetologica* **54**, 279–298 (1998).
10. Avise, J. *Phylogeography: The History and Formation of Species* (Harvard Univ. Press, Cambridge, MA, 2000).
11. Riddle, B. R., Hafner, D. J., Alexander, L. F. & Jaeger, J. R. *Proc. Natl Acad. Sci. USA* **97**, 14438–14443 (2000).
12. Orr, M. R. & Smith, T. B. *Trends Ecol. Evol.* **13**, 502–506 (1998).
13. Schneider, C. J., Smith, T. B., Larison, B. & Moritz, C. *Proc. Natl Acad. Sci. USA* **96**, 13869–13873 (1999).
14. Rice, W. R. in *Endless Forms: Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 261–270 (Oxford Univ. Press, New York, 1998).
15. Gray, D. A. & Cade, W. H. *Proc. Natl Acad. Sci. USA* **97**, 14449–14454 (2000).

Water structure

Order and oddities

Srikanth Sastry

Water is a common but unusual liquid. Precise measures of the arrangement of molecules in water may help us to better understand some of its peculiar properties.

The structure of matter on microscopic scales is a pervasive theme in modern science. The successes of solid-state physics are built on exploiting the regularity of atomic arrangements in crystals. Chemists and materials scientists focus on relationships between the structure and properties of molecules and materials. Molecular biologists are concerned with the interplay of structure and function in bio-

logical molecules and their assemblies. Describing the crystalline order of solids is relatively straightforward, but the order found within 'disordered' materials, such as a liquid, glass or a protein molecule, is far more subtle and trickier to capture. On page 318 of this issue, Errington and DeBenedetti¹ show that combining two measures of the structural order in water — the most familiar of liquids but one that is notoriously peculiar

— reveals striking patterns of structural change, and also its relation to water's unusual properties.

Perhaps the best-known peculiarity of water is its 'density maximum' at 4 °C (at atmospheric pressure); cooling or heating water from this temperature causes it to become less dense. An equally striking but less-known anomaly is that, as the density of water is increased, water molecules begin to move about — or diffuse — faster, but only up to a point known as the 'diffusivity maximum'. At higher densities, the diffusivity decreases with increasing density, just as in normal liquids. By using computer simulations, Errington and DeBenedetti identify a 'structurally anomalous' region in the density–temperature phase diagram of water, in which order decreases with increasing density. They also find that all the unusual density and diffusivity behaviour occurs in this region, suggesting a precise link between anomalies in structure and behaviour.

In crystalline solids, atoms or molecules are arranged in a regular array. No such obvious order exists in liquids. A microscopic 'snapshot' of a liquid would reveal atoms in a disordered jumble, much like people packed during rush hour into a crowded subway in New York, or, even better, a local train in Mumbai. There is some form of order nevertheless, and it arises from what the English humorist P. G. Wodehouse describes as the law that "a given spot... at a given moment of time can be occupied by only one body". From where one passenger stands, another passenger in any direction must be found a well-defined distance away, and the next at roughly twice that distance, and so on. But such order doesn't persist over long distances.

Neuroscience

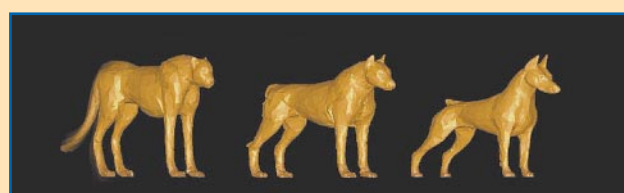
Cats, dogs and categories

Any child can tell a cat from a dog. But the difference has to be learned, and describing it is far from simple. Cats include cheetahs, lions and tabbies; dogs include Siberian huskies and dachshunds. How do we make the jump from recognizing a particular set of features to establishing a more general concept or category that will help us extrapolate to new situations?

Writing in last week's *Science* (291, 312–315; 2001), David J. Freedman and colleagues describe how they have explored this question by training monkeys to distinguish between 'catness' and 'dogness'. The authors used computer graphics to create blended images from a set of three dog and three cat images. An example of a cheetah, a Doberman and, in between, a blend of the two is

shown on the right. They then taught the monkeys to indicate, by releasing a lever, whether a sample image was of the same type as a test cat or a test dog. Monkeys, it turns out, are good at learning this distinction: even when the image was 60% cat and 40% dog, the monkeys reliably reported that it was like a cat. Furthermore, monkeys were not simply memorizing specific blends of cats and dogs as belonging to one category, because new blends were tested during the experiment.

To find out how these categories are represented in the brain, the authors recorded neural activity in the lateral prefrontal cortex — an area of the frontal lobes previously implicated in guiding complex behaviours — while the monkeys performed the task. Surprisingly, they found category



information represented at the level of single neurons. That is, regardless of whether the image was 60%, 80% or 100% dog, individual neurons responded in a similar way; but they responded differently for 60%, 80% or 100% cat.

Obviously, these category representations were the result of training — neurons in a monkey's lateral prefrontal cortex probably don't care about 'dogness' under normal circumstances. Indeed, the authors

went on to train one of their monkeys on a new, more abstract categorization of the same images, and showed that neurons no longer distinguished cats and dogs as they did previously, but now coded for the new categories. How these representations come to be formed rapidly and reversibly in this part of the brain is not going to be easy to answer. But it is clearly closely related to how we learn to categorize our world into meaningful concepts.

Hemai Parthasarathy

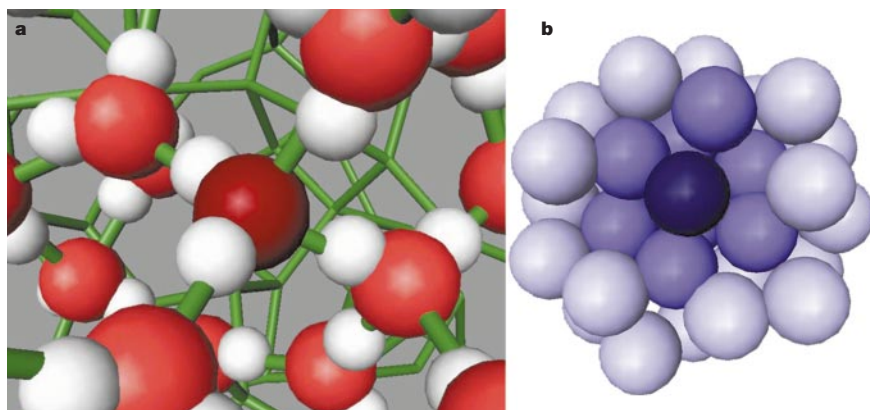


Figure 1 Molecular structure in liquids. **a**, In water, molecules typically form hydrogen bonds (green bars) with four neighbours, and the liquid displays an open, loosely packed, structure. Hydrogen atoms are shown in white, oxygen atoms are red, and the central oxygen atom is dark red. Only the molecules within the first two neighbouring shells are shown, but all the bonds are indicated. **b**, In a simple atomic liquid, the packing is much more dense. The first two neighbouring shells of the dark-blue central atom are shown, and the atoms in front of the plane of the central atom have been removed, for clarity.

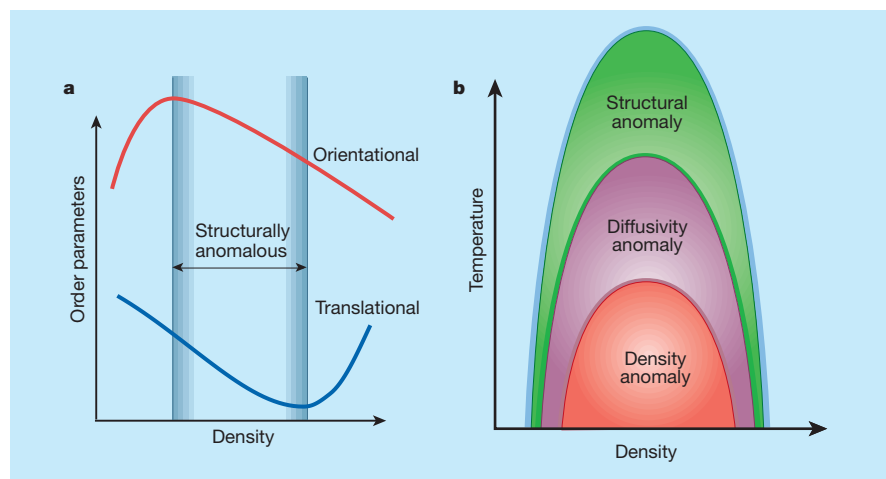


Figure 2 A possible link between structural order and anomalous behaviour in liquid water. **a**, The typical density dependence of orientational and translational order parameters. Errington and Debenedetti¹ identify a 'structurally anomalous' region, bounded at low densities by an orientational-order maximum, and at high densities by a translational-order minimum. **b**, In the density-temperature phase diagram, the 'structurally anomalous' region forms a dome, which encompasses the regions of dynamic (diffusivity) and thermodynamic (density) anomalies.

Whereas repulsions determine this short-range order in normal liquids², the architects of local structure in water are hydrogen bonds. These bonds require pairs of molecules not only to be optimally separated, but also to have a specific orientation. In the most energetically favoured structure, each molecule has four hydrogen-bonded neighbours that sit at the corners of a regular tetrahedron (Fig. 1a), in contrast to about twelve neighbours in an atomic liquid (Fig. 1b). Water molecules are therefore less densely packed, and water possesses a far greater degree of local 'orientational' order than normal liquids. Qualitatively, many of water's peculiarities can be understood from this observation alone³.

Errington and Debenedetti shed new light on the relationship between the structure of water and its anomalous behaviour

by focusing on global aspects of structural change. They define two order parameters. The first quantifies translational order: how neighbours of a molecule are distributed, on average, in their distance from the molecule. The second parameter quantifies orientational order: the extent to which neighbouring molecules are at specific angles with respect to each other. In normal liquids, both translational and orientational order increase with density. In water, however, for a range of densities and temperatures, both translational and orientational order decrease when density increases, because compression leads to a disruption of the network of hydrogen bonds. Errington and Debenedetti identify this as their 'structurally anomalous' region. At a fixed temperature, this region is bounded by an orientational-order maximum at the low-

Daedalus

David Jones

David Jones, author of the Daedalus column, is indisposed.

density end, and by a translational-order minimum at the high-density end (Fig. 2a). The structurally anomalous region forms a 'dome' in the density-temperature diagram (Fig. 2b).

Errington and Debenedetti show that the regions of peculiar diffusivity and density are nested like Russian dolls within the region of structural anomaly (Fig. 2b). So liquid water appears to have a hierarchy of anomalies: the structural anomalies occur over the broadest range of temperatures and densities, within which are found both the diffusivity and density anomalies.

Another curious pattern emerges when the values of the orientational and translational order parameters, at various densities and temperatures, are plotted against each other in a 'parametric plot'. All available pairs of values fall to one side of a boundary line and, remarkably, points corresponding to the structurally anomalous region all appear to fall on the boundary itself (see Fig. 5 on page 320). Because the boundary is a one-dimensional line, the translational order and orientational order are strongly coupled within the structurally anomalous region.

Errington and Debenedetti's observations raise interesting questions and open a new line of investigation. The characterization of structural anomaly in terms of the strong coupling between translational order and orientational order may help to identify precise conditions necessary for anomalous behaviour. But at present it isn't clear why this observed relationship and the nested pattern of structural, dynamic and thermodynamic anomalies hold, and whether we should expect to find them in other liquids as well.

It will be useful to apply the present analysis to other network-forming liquids that display structural and other anomalies — silicon and silica, which form tetrahedral networks in the liquid and solid states, are obvious candidates. If these results prove robust, they will have established a significant link between anomalous behaviour and the strongly correlated changes of translational and orientational order in network-forming liquids.

Srikanth Sastry is at the Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur Campus, Bangalore 560064, India.

e-mail: sastry@jncasr.ac.in

1. Errington, J. R. & Debenedetti, P. G. *Nature* **409**, 318–321 (2001).
2. Chandler, D., Weeks, J. D. & Andersen, H. C. *Science* **220**, 787–794 (1983).
3. Ball, P. *H₂O: A Biography of Water* (Weidenfeld and Nicolson, London, 1999).

Obituary

Louis Néel (1904–2000)

Louis Néel, who died on 14 November 2000 at the age of 95, was a dominant presence in French physics after the Second World War. By the time he retired in the mid-1970s, he had transformed Grenoble from a sleepy provincial backwater into a European centre for scientific research. Grenoble is now endowed with a huge nuclear centre, a first-rate technical university, a suite of specialized laboratories for the Centre Nationale de la Recherche Scientifique, a high-flux neutron reactor, and a European synchrotron radiation source. All of these, and more, Néel founded, directed or promoted.

Unusually for a Frenchman of his status, he made his entire career outside Paris, apart from four years at the Ecole Normale Supérieure, where he emerged top of the 1928 class in physical sciences.

The son of a government official, Néel spent a peripatetic childhood following his father's postings across the French provinces and North Africa, before being sent back to his native Lyons to prepare for entry to the Ecole Normale. Attracted to the study of magnetism, in 1928 he accepted an offer of an assistantship with Pierre Weiss in Strasbourg. Already present in Néel's doctoral thesis was the idea of extending his mentor's molecular field theory of ferromagnetism to understand the constant paramagnetism of certain metals whose magnetic properties were almost independent of temperature. This led him to predict the existence of antiferromagnetism — examples of which were discovered several years later — and thence to the theory of ferrimagnetism. Ferrimagnetic oxides with the Néel structure of two unequal and oppositely aligned magnetic sublattices are now used in most permanent magnets, recording media and high-frequency magnetic materials. They include the archetypal magnet, lodestone.

Néel's forte was phenomenological theory — understanding complex magnetic phenomena in terms of simple and solvable models that allowed 'back of the envelope' calculations. His work contained no profound insights into the origin of magnetism; the fundamental physics was already in place after the 1930 Solvay conference. Néel sought an understanding of the phenomena in broad physical terms rather than in mathematical detail, preferring to "explore virgin forests rather than to cultivate the vicarage garden".

At heart he was a craftsman and no follower of fashion. Once asked by Anatole



Father of Grenoble's scientific fortune

Abraham why he had persisted with his theory of antiferromagnetism in the light of Lev Landau's prediction that only ferromagnetism was possible in nature, Néel thanked heaven he was not that smart. He enjoyed concentrating on practical issues: degaussing ships to protect them from destruction by magnetic mines; and explaining the natural magnetism of rocks in terms of the properties of oxide nanoparticles. The consequences were far-reaching. They included the establishment of the theory of global plate tectonics based on palaeomagnetic data, development of stealth bombers and the foundation of the modern magnetic-recording industry. His name is associated with a dozen concepts found in the toolkits of magnetic engineers from California to Kyoto — the Néel point, Néel structures, Néel walls, Néel's theory of superparamagnetism, induced anisotropy, exchange bias, orange-peel coupling... As a fellow Nobel laureate J. H. Van Vleck remarked in the preface to the English edition of *The Selected Works of Louis Néel* (Gordon & Breach, New York, 1988): "The name of the Place de l'Etoile may be changed, but Curie point and Néel point will forever belong in the terminology of physics."

Néel was the last of a generation of scientists who could afford to write almost entirely in French, in a sparse style with short numbered paragraphs and a disregard for units which belied a sure sense of the physical magnitudes involved.

Nevertheless, his ideas found their way first to Paris, where he was elected to the Academy of Sciences in 1952, and then to Stockholm — he shared the 1970 Nobel Prize in Physics with Hannes Alfvén.

Néel published a volume of memoirs *Un Siècle de Physique* (Odile Jacob, Paris, 1991), which offers a characteristically astute assessment of his own achievements, as well as of the foreign colleagues whom he encountered on his limited travels, and on the NATO Scientific Committee where he represented France for 20 years. Much of his book chronicles the transformation of the scientific environment in Grenoble, where he found a welcome after the Franco-German armistice in 1940. There he gathered and animated a circle of co-workers in the interpenetrating institutes with impenetrable acronyms that constitute the research and university scene in that city on the western edge of the Alps. When the Laboratoire d'Electrostatique et de Physique du Métal that Néel founded in 1945 was reconstituted as five separate laboratories in the Grenoble polygon in 1971, it was carrying out a range of work that could hardly have been divined from its original name. He foresaw the importance of large-scale facilities for condensed-matter research, not only for physics but also for biology, and ensured that Grenoble was endowed with world-class facilities for high magnetic fields, neutron scattering and synchrotron radiation. Franco-German collaboration was the key to realizing these ventures, which are a microcosm of the process of European integration. All of these institutes and the craze for winter sports have secured Grenoble's fortune.

Sensitive to the organic relation between applied science and technology, and the different perspectives of the researcher and the industrialist, there was less Néel could achieve in this area. He held patents on nanoparticulate permanent magnets and thin-film magnetic memory, but the spirit of the times frowned on the use of public resources for the benefit of private industry. Entrepreneurship was to arrive later.

Néel's success as a scientist rested on the simplicity and applicability of his ideas, and his physical insight. His success as a manager rested on a universal respect for his integrity and fairness. Never eloquent, he was thoroughly professional in all his dealings and a dogged champion of his vision for Grenoble, remaining admirably lucid and alert to the end of his long life.

Michael Coey

Michael Coey is in the Department of Physics, Trinity College Dublin, Dublin 2, Ireland.
e-mail: jcoey@tcd.ie

A viable herd of genetically uniform cattle

Deleterious alleles seem to have been purged in a feral strain of inbred cows.

Inbreeding, which can lead to the loss of genetic variation or the accumulation of deleterious alleles, has been shown to reduce fitness in wild¹, zoo, laboratory² and farmed³ animals. But it has been proposed that when combined with selection, inbreeding may purge deleterious alleles⁴. Here we provide support for this hypothesis in a study of the Chillingham cattle, which shows that this viable herd is almost genetically uniform. The homozygosity of this herd far exceeds that of other cattle⁵ and that found in wild populations of other mammalian species^{6,7}.

This feral herd, which lives in a park in northern England, is thought to have experienced no immigration for at least 300 years (Fig. 1). Despite this genetic isolation, records of calvings and deaths suggest that there has been no drop in fertility or viability⁸. In 1947 the population crashed to five bulls and eight cows; as of 30 October 2000 it numbered 49 individuals. Studies of blood groups and biochemical polymorphisms that represent a small number of genetic loci showed homozygosity for the same alleles at each locus⁸.

We obtained tissues from calves, adult cows over 6 years old and bulls that died in 1998–99 ($n = 13$). Family relationships and causes of death were unknown; there was no evidence of infectious disease. DNA from the samples was scored for 25 microsatellite markers. These markers are highly polymorphic in cattle⁵, with a heterozygosity of typically 70%, and cover 15 of the 29 autosomes. Three markers are located around the bovine major histocompatibility complex (MHC) region, where selection might maintain polymorphism⁹.

We successfully amplified between 3 and 23 markers for each of the 13 samples, obtaining a total of 225 marker genotypes. (The variation was due to the poor quality of DNA extracted from some hair-root samples.) All samples showed an identical homozygous genotype for 24 of 25 markers. For one marker (*HUJ616* on chromosome 13, amplified or sequenced from 11 samples), nine individuals were heterozygous for the same alleles, and two were homozygous for the same allele, with one allele in common with the heterozygous individuals.

We believe that this single heterozygous marker represents the persistence of ancestral heterozygosity. Assuming 67 generations since 1700, a coefficient of variation of 50% in the population size, and a mean number of sires and dams per generation of 3 and 15, respectively, the effective population size N_e (ref. 10) averages 8.0, and the



Figure 1 Chillingham cattle (*Bos taurus*). This herd, which now comprises 49 animals, lives in isolation in a park in the north of England. The strain has been studied extensively, including by Darwin, because it has remained viable and fertile despite at least 300 years of total inbreeding.

expected proportion of remaining ancestral heterozygosity is $[1 - 1/(2N_e)]^{67} = 0.013$. The probability of segregation at a multi-allelic locus with ancestral heterozygosity of 0.70 after 67 generations is approximately $3 \times 0.70 \times [1 - 1/(2N_e)]^{67} = 0.028$ (ref. 11). Thus the probability of finding one heterozygous marker out of 25 markers sampled, using the binomial distribution and assuming unlinked loci, is 0.35, and our data are consistent with the population having been closed for 300 years.

The observed genotype frequencies at the segregating locus are not significantly different from Hardy–Weinberg equilibrium ($P = 0.057$, exact test). Re-typing, re-scoring, and sequencing the single marker *HUJ616* gave the same results, making a procedural artefact or a recent mutation unlikely. Heterozygosity could be maintained by selection. Homozygosity was observed at two linked markers on chromosome 13 (9 and 5 animals typed successfully at each). But these markers were 5 and 3 centimorgans from *HUJ616*, respectively, too distant to observe a remaining ancestral heterozygous segment.

For the bovine genome (which has a total map distance of 30 morgans — the expected number of crossovers in the genome during meiosis is 30), with random breeding in isolation for 67 generations, initial complete heterozygosity is predicted to result in 30 heterozygous segments with a mean length of only 1.5 centimorgans¹². We cannot determine from our data the most likely cause of observed heterozygosity, but

it is consistent with the segregation of alleles at a neutral locus in a small population which has been closed for 300 years.

We believe the continuing viability of the herd shows that deleterious alleles have been purged. No chromosome segments appear heterozygous, apart from a single marker. There is no evidence of heterozygosity near the MHC complex, implying that selection by disease or parasites has not been important at this locus, or perhaps that an optimum haplotype has been fixed. The homozygosity of the Chillingham herd might help to elucidate the bovine genome and the genetics of disease resistance¹³.

P. M. Visscher*, **D. Smith†**, **S. J. G. Hall‡**, **J. L. Williams†**

*Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK
e-mail: peter.visscher@ed.ac.uk

†Roslin Institute, Roslin EH25 9PS, UK

‡School of Agriculture, De Montfort University, Caythorpe Court, Grantham NG32 3EP, UK

1. Saccheri, I. *et al.* *Nature* **392**, 491–494 (1998).
2. Frankham, R. & Ralls, K. *Nature* **392**, 441–442 (1998).
3. Wiener, G., Lee, G. J. & Woolliams, J. A. *Anim. Prod.* **55**, 89–99 (1992).
4. Lacy, R. C. & Ballou, J. D. *Evolution* **52**, 900–909 (1998).
5. MacHugh, D. E., Loftus, R. T., Bradley, D. G., Sharp, P. M. & Cunningham, E. P. *Proc. R. Soc. Lond. B* **256**, 25–31 (1994).
6. Manotti-Raymond, M. A. & O'Brien, S. J. *J. Hered.* **86**, 319–322 (1995).
7. Goodman, S. J. *Mol. Biol. Evol.* **15**, 104–118 (1998).
8. Hall, S. J. G. & Hall, J. G. *J. Zool.* **216**, 479–493 (1988).
9. Paterson, S., Wilson, K. & Pemberton, J. M. *Proc. Natl Acad. Sci. USA* **95**, 3714–3719 (1998).
10. Caballero, A. *Heredity* **73**, 657–679 (1994).
11. Kimura, M. *Proc. Natl Acad. Sci. USA* **41**, 144–150 (1955).
12. Stam, P. *Genet. Res.* **35**, 131–155 (1980).
13. Anderson, C. *Nature* **355**, 382 (1992).

Surface science

Topology of two-dimensional C₆₀ domains

Two-dimensional systems possess a unique topological ordering that is not found in either three- or one-dimensional systems¹. Using high-resolution scanning tunnelling microscopy, we show here that a 60-carbon-atom (C₆₀) array on a self-assembled monolayer of an alkythiol forms an ideal two-dimensional system which has another novel topological order originating from the orientational degrees of freedom. At a temperature of 5 K, the two-dimensional C₆₀ forms a domain structure in which the correlation function of the molecular orientation within a domain is constant anywhere (so every C₆₀ has the same orientation) but changes abruptly at domain boundaries. Remarkably, the positional order and the bond-orientational order are both fully preserved across domain boundaries.

The self-assembled monolayer substrate was prepared using a standard procedure² and transferred into an ultra-high-vacuum and low-temperature scanning tunnelling microscope (STM) chamber where a submonolayer of C₆₀ was then thermally evaporated onto the substrate. The STM images show that the C₆₀ molecules form close-packed hexagonal arrays, with a nearest-neighbour distance of 10 Å (ref. 3). At room temperature, molecules at the edge of an array can detach readily and diffuse to another part of the same array or to other nearby arrays. Each C₆₀ displays a smooth hemispherical protrusion, suggesting that the C₆₀ molecules are rotating freely at this temperature. This contrasts with C₆₀ adsorbed on metal or semiconductor surfaces, when the molecular rotation is frozen even at room temperature owing to strong binding of C₆₀ on the substrate⁴.

At 77 K, the C₆₀ molecule appears as a hemisphere, a tilted doughnut, or an asymmetric dumb-bell, each being consistent with a rotating pattern around a fixed axis. When the sample is cooled down to 5 K, all C₆₀ molecules in the STM image (Fig. 1a) start to reveal an identical internal fine structure that closely matches the well-known cage structure (stick model in Fig. 1b). To our knowledge, this is the first time that the C₆₀ native cage structure has been seen in the STM.

From the evolution of C₆₀ internal patterns with decreasing temperature, we conclude that there is an ordered rotational motion of the two-dimensional C₆₀ array, as in the case of bulk C₆₀, except that the two-dimensional rotationally ordered phase persists down to 77 K (already 13 K below the bulk freezing temperature³). Unlike the orientation-related glassy phase in bulk

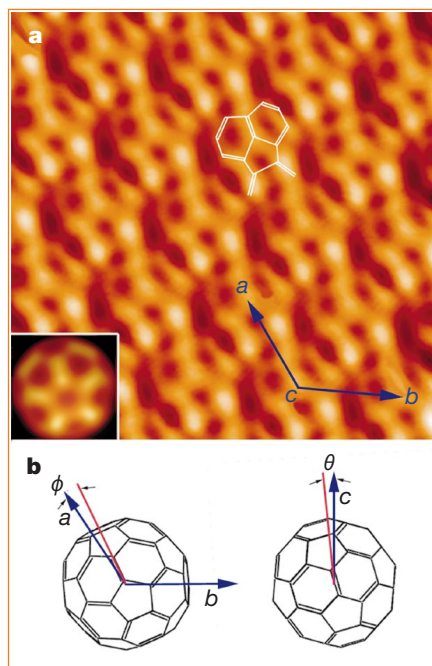


Figure 1 The native cage structure of C₆₀ molecules, seen using a scanning tunnelling microscope (STM). **a**, STM image (35 × 35 Å) of a C₆₀ lattice taken at 5 K with −2.0 V sample bias. Detailed internal features of the C₆₀ molecule are evident that closely resemble the C₆₀ cage structure and match the theoretical simulation shown in the inset. **b**, Top view (left) and side view (right) of a stick model outlining the C₆₀ orientations obtained by simulation. The *c*-axis points out of the image.

C₆₀, these two-dimensional C₆₀ arrays form orientationally ordered domains at 5 K. These observations, together with the fact that the C₆₀ molecules are not in an epitaxial (or commensurate) arrangement with respect to the self-assembled monolayer (the lattice of the underlying monolayer has a rectangular unit cell of 9.994 × 8.655 Å, containing four alkythiol molecules; see ref. 5, for example), are indicative of a very weak interaction between the C₆₀ and the alkythiol.

Figure 1a (inset) shows a simulated STM image of a C₆₀ structure obtained by integrating the electron density of states on a C₆₀ from the Fermi level to the bias voltage⁶. The C₆₀ orientation is tuned for the best agreement between the simulation and the experiment. The final molecule orientation has an edge atom facing towards the substrate and can be specified by the azimuthal ($\theta = 1.5^\circ$) and polar ($\phi = 0.6^\circ$) rotation, as defined in Fig. 1b. This remarkable conformity between simulation and experiment allows us confidently to identify C₆₀ domains with different molecular orientations.

Although most of the C₆₀ arrays feature a single orientation, we have also observed a single array consisting of domains of two different orientations (Fig. 2a), in which the boundary separates two distinguishable domains by their internal features. Comparison with our simulation results enabled

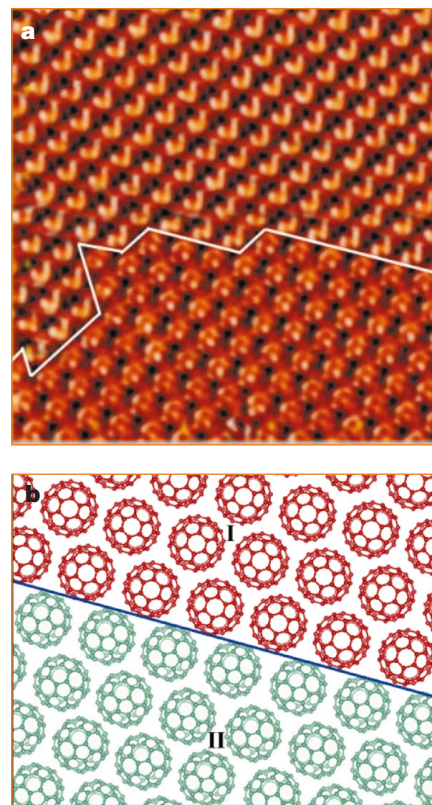


Figure 2 Domain boundary of a two-dimensional C₆₀ array. **a**, STM image (100 × 100 Å) of two molecular orientations and the domain boundaries. Both the positional order and bond-orientational order are fully preserved and no defect exists along the domain boundaries. **b**, Representation of the two molecular orientations derived by comparison with the theoretically simulated images.

us to identify these two orientations as those shown in Fig. 2b. Apart from the two different molecular orientations, there is no positional defect at the domain boundary and the entire C₆₀ array maintains the perfect translation symmetry for the centres of the C₆₀. Although the two orientation domains are quite stable, we found that the domain boundary fluctuates in time, especially when the tip is brought closer to the surface. No domain boundaries are found in the adjacent self-assembled monolayer substrate that can be correlated to the C₆₀ domains.

A C₆₀ molecule consists of 60 single and 30 double carbon bonds. There is a small charge deficiency between the single bonds and a charge excess on the double bonds. Thus the intermolecular potential contains, in addition to the van der Waals potential, a small bond-to-bond Coulomb energy dependent on relative molecule orientations⁷. The latter results in a variation of the domain energy by about 0.1 eV per molecule, roughly one tenth of the binding energy of a C₆₀ in a two-dimensional close-packed array (Y.J. *et al.*, unpublished results).

This large energy-scale difference is probably responsible for the preservation of the positional and bond-orientational

order across the domain boundary, even though most of the boundary structures yield positive boundary energy of around $0.01 \text{ eV } \text{\AA}^{-1}$. The fact that C_{60} forms orientationally ordered domains suggests that the magnitude of the domain energy variation sets the upper bound of the rotational barrier due to the monolayer substrate. However, an absence of positional and bond-orientational domain walls, such as those in noble gases adsorbed on graphite⁸, implies that the lateral variation of the substrate fields is much smaller than the C_{60} intermolecular potential. We conclude that the novel topological order observed here must be an intrinsic property of a two-dimensional system.

J. G. Hou^{*†}, Yang Jinlong^{*†},
Wang Haiqian^{*}, Li Qunxiang[†],
Zeng Changgan^{*}, Yuan Lanfeng[†],
Wang Bing^{*}, D. M. Chen^{*†}, Zhu Qingshi[†]

^{*}Structure Research Laboratory and [†]Open Laboratory of Bond Selective Chemistry, University of Science and Technology of China, Hefei 230026, PR China

e-mail: jghou@ustc.edu.cn

[‡]The Rowland Institute for Science, Cambridge, Massachusetts 02142, USA

1. Sinha, S. K. (ed.) *Ordering in Two Dimensions* (Elsevier North Holland, Amsterdam, 1980).
2. Evans, S. D. & Ulman, A. *Chem. Phys. Lett.* **170**, 462–466 (1990).
3. Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego, 1996).
4. Sakurai, T. et al. *Prog. Surf. Sci.* **51**, 266–408 (1996).
5. Fenter, P. in *Thin Films — Self-Assembled Monolayers of Thiols* Vol. 24 (ed. Ulman, A.) 121–125 (Academic, San Diego, 1998).
6. Hou, J. G. et al. *Phys. Rev. Lett.* **83**, 3001–3004 (1999).
7. Lu, J. P., Li, X.-P. & Martin, R. M. *Phys. Rev. Lett.* **68**, 1551–1554 (1992).
8. Birgeneau, R. & Horn, P. M. *Science* **232**, 329–336 (1986).

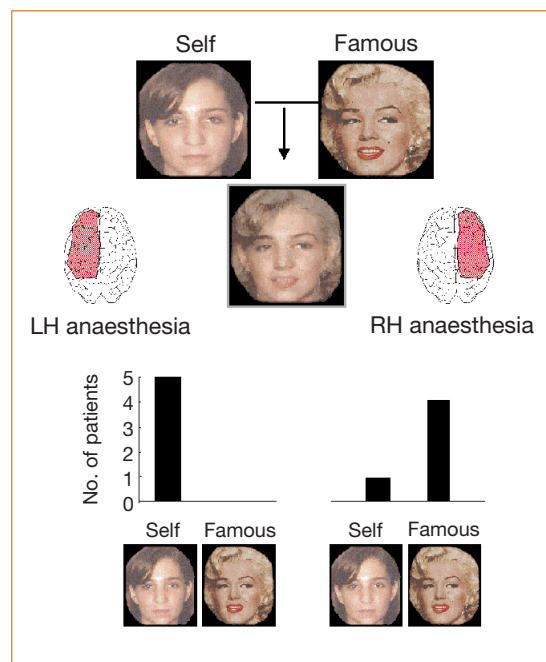
Neurology

Self-recognition and the right hemisphere

Although monkeys can perceive complex stimuli such as faces¹, only the higher apes are capable of recognizing their own face in a mirror². Here we show that in humans the right hemisphere of the brain seems to be preferentially involved in self-face recognition. Our findings indicate that neural substrates of the right hemisphere may selectively participate in processes linked to self-awareness.

We first studied a group of patients undergoing the intracarotid amobarbital (Wada) test. The Wada test involves anaesthetization (that is, inactivation) of one cerebral hemisphere in order to provide information regarding cerebral dominance for language and other cognitive phenomena. Five right-handed (left hemisphere is language-dominant) patients, who were undergoing Wada tests for evaluation for surgery to treat epilepsy, were shown pic-

Figure 1 Five patients (a real patient is not shown) were presented with a picture showing a morph of a face that was composed of their own face and a famous face during the time when either the right or the left hemisphere of their brain was anaesthetized. Following anaesthesia of the left hemisphere (LH), patients selected the 'self' face as having been shown to them (5/5); after anaesthesia of the right hemisphere (RH), patients selected the famous face as the one they had viewed (4/5). In a second experiment with 10 normal subjects (results not shown), transcranial magnetic stimulation was delivered to the motor cortex of the RH or LH during self-famous or familiar-famous morph display. The amplitude of the resulting motor-evoked potentials (MEPs) was significantly greater for the RH than for the LH during presentation of self morphs ($M = 1.26 \text{ mV}$ and $M = 1.02 \text{ mV}$; s.e., 0.09, respectively) and the former (RH, self morph) was also significantly greater than the MEP amplitude from both hemispheres during presentation of familiar morphs (RH, familiar: $M = 1.04 \text{ mV}$, s.e., 0.07; LH, familiar: $M = 1.03 \text{ mV}$, s.e., 0.08).



tures of faces generated by morphing the picture of a famous person with the patient's own face (Fig. 1). Patients were instructed to remember the picture presented. Different pictures were presented during selective anaesthesia of the right and the left hemispheres.

After recovery from anaesthesia, patients were given a forced-choice task in which they had to choose the picture of the face that they had been shown. The two choices were the pictures from which the morphed image had been generated (self and famous), although neither choice had actually been presented during anaesthesia. Following anaesthesia of the left hemisphere, all five patients selected the 'self' face as the one they thought had been presented; however, after anaesthesia of the right hemisphere, four out of the five selected the famous face. These results suggest that the anterior right hemisphere may be critically engaged in detecting the self face.

To find out whether a similar effect operates in normal subjects, we presented morphed photographs to ten volunteers (instructed to attend to the photographs) and delivered transcranial magnetic stimulation to the motor cortex of the left or right hemisphere at random times (100–450 ms) during picture presentation. We measured the amplitude of the motor-evoked potentials induced by this stimulation in the contralateral first dorsal interosseous muscle as an indicator of the amount of activation of each hemisphere³ during exposure to pictures morphing their own or a familiar face with that of a famous person.

We found that there was a significant association ($P < 0.05$) between subjects' hemisphere activation and their exposure to the 'self' and 'familiar' conditions. Post-hoc

Bonferroni tests revealed that motor-evoked potentials were significantly greater from the right hemisphere while subjects viewed pictures containing elements of their own face than for all other conditions.

Some people can suffer from neglect or misidentification of their own extremities (a condition known as asomatopagnosia) after damage⁴ to or anaesthetization⁵ of the right hemisphere. It is also known that patients with lesions to the right fronto-temporal cortex may experience a cognitive detachment from self⁶. This recently evolved network includes prefrontal regions of the brain and demonstrates a high degree of lateralization⁷ in humans and apes, but not in monkeys. It is conceivable that a right-hemisphere network gives rise to self-awareness, which may be a hallmark of higher-order consciousness⁸.

Julian Paul Keenan, Aaron Nelson*, Margaret O'Connor, Alvaro Pascual-Leone
Behavioral Neurology Unit, Department of Neurology, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA

e-mail: jkeenan@caregroup.harvard.edu

*Division of Cognitive and Behavioral Neurology, Department of Behavioral Neurology, Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA

1. Pineda, J. A., Sebesteny, G. & Nava, C. *Brain Res. Cogn. Brain Res.* **2**, 1–12 (1994).
2. Gallup, G. G. *Science* **167**, 86–87 (1970).
3. Tormos, J. M. et al. *Neurology* **49**, 487–491 (1997).
4. Feinberg, T., Haber, L. & Leeds, N. *Neurology* **40**, 1391–1394 (1990).
5. Meador, K., Loring, D., Feinberg, T., Lee, G. & Nichols, M. E. *Neurology* **55**, 816–820 (2000).
6. Wheeler, M. A., Stuss, D. & Tulving, E. *Psychol. Bull.* **121**, 331–354 (1997).
7. LeMay, M. & Geschwind, N. *Brain Behav. Evol.* **11**, 48–52 (1975).
8. Sperry, R., Zaidel, E. & Zaidel, D. *Neuropsychologia* **17**, 153–166 (1979).

Geophysics

Life, geology and snowball Earth

According to the 'snowball Earth' hypothesis, a series of global glaciations occurred 750–580 million years ago, each lasting for millions of years and ending in a scorching heat caused by an extreme enrichment of atmospheric greenhouse gases. Hyde *et al.*¹ have used climate models to simulate this global glaciation, finding in one case an alternative climate scenario in which a partially frozen Earth has ice-free oceans equatorward of 25° latitude. We do not believe that this 'slushball' Earth is consistent with the most striking geological and palaeomagnetic observations explained by the snowball Earth hypothesis.

Palaeomagnetic and geological data from Neoproterozoic glacial deposits indicate that glaciations were long-lived (lasting for millions of years)^{2,3} and locally associated with iron formations. The glacial deposits are covered by extraordinary sequences of carbonate sediments called 'cap' carbonates, which have unusual textures and low $\delta^{13}\text{C}$ values. The snowball Earth hypothesis can explain all these observations³, whereas a semifrozen (slushball) Earth does not.

In the snowball Earth hypothesis, a runaway ice–albedo feedback leads to a planet frozen to the Equator⁴. Extremely large increases in carbon dioxide are required to terminate the glaciation and overcome the climate stability imposed by the high planetary albedo⁵. In Hyde *et al.*'s model¹, with all the continents covered in ice, volcanic emissions without chemical weathering would cause atmospheric CO_2 levels to rise. But with ice-free tropical oceans, even a modest rise in CO_2 would cause the tropical glaciation to be short-lived. The exact duration would depend on the extent of chemical equilibration between sea water and calcium carbonate in the deep ocean, but it would be far less than the roughly 10 million years estimated from analysis of basin subsidence³. It is also unclear how iron could accumulate in sea water to produce banded iron formations if tropical oceans were ice-free. Moreover, the model of Hyde *et al.* predicts a progressive deglaciation from the Equator to the poles occurring at much lower CO_2 concentrations, no higher than during the Cretaceous period.

Support for extreme increases in CO_2 in the aftermath of the glaciations comes from cap carbonates, which are nearly ubiquitous features of Neoproterozoic glacial deposits. These carbonate rocks have distinctive features, including "knife-sharp" contacts with glacial deposits and unusual textures indicative of rapid precipitation on the sea floor⁶. The cap carbonates, which have been singled out as a climate paradox of Neo-

proterozoic geology⁷, are predicted by the snowball Earth hypothesis to be a consequence of intense carbonate and silicate weathering in the aftermath of the deglaciation. It is hard to reconcile the global occurrence of such a pulse of intense carbonate precipitation immediately after the termination of the ice ages with the progressive retreat of ice at moderate CO_2 levels predicted by the Hyde *et al.* model.

Excitement over whether a semifrozen Earth might explain the geological observations stems from concern for the survival of eukaryotic life in such extreme and extended glaciations⁸. The critical feature is the survival of groups of photosynthetic algae that evolved before the glaciations. The survival of metazoans, as discussed by Hyde *et al.*, is less problematic because such organisms (if they existed) could live wherever primary producers (photosynthetic or chemosynthetic) were still active. Photosynthetic algae could survive a series of glaciations in refugia near volcanic islands, such as Iceland and Hawaii, or beneath thin equatorial ice cover⁹. Evolution might well be stimulated by this prolonged genetic isolation, and by perturbations of biogeochemical cycles during the postglacial, ultra-greenhouse climate. This is consistent not merely with the survival of eukaryotic life, but also with the coincident radiation of metazoa and other groups⁸.

Daniel P. Schrag, Paul F. Hoffman

Department of Earth and Planetary Sciences,
Harvard University, 20 Oxford Street, Cambridge,
Massachusetts 02138, USA

e-mail: schrag@eps.harvard.edu

- Hyde, W. T., Crowley, T. J., Baum, S. K. & Peltier, R. *Nature* **405**, 425–429 (2000).
- Sohl, L. E., Christie-Blick, N. & Kent, D. V. *Geol. Soc. Am. Bull.* **111**, 1120–1139 (1999).
- Hoffman, P. F., Kaufman, J. A., Halverson, G. P. & Schrag, D. P. *Science* **281**, 1342–1346 (1998).
- Budyko, M. I. *Tellus* **21**, 611–619 (1969).
- Caldeira, K. & Kasting, J. F. *Nature* **359**, 226–228 (1992).
- Kennedy, M. J. *J. Sedim. Res.* **66**, 1050–1064 (1996).
- Fairchild, I. J. in *Sedimentology Review 1* (ed. Wright, V. P.) 1–16 (Blackwell, Oxford, 1993).
- Runnegar, B. *Nature* **405**, 403–404 (2000).
- McKay, C. P. *Geophys. Res. Lett.* **27**, 2153–2156 (2000).

Hyde et al. reply — We are not convinced that the data discussed by Schrag and Hoffman can be interpreted in only one way. With respect to the duration of glaciation, calculations¹ suggest that the open-water solution could have persisted up to CO_2 levels of about four times those at present. With buffering from the reactive ocean carbonate reservoir, the time required for degassing to raise atmospheric concentrations to this level would be about 0.2–0.6 million years (L. A. Derry, personal communication). As ice sheets would still discharge ground carbonate into the ocean, the buffering time would be extended. This timescale is consistent with range estimates derived from palaeomagnetic data² and geochemical calculations³.

Our open-water results also agree with

the environment of deposition of some glaciomarine sediments⁴. Iron is not a consistent feature of Neoproterozoic glacial sequences⁵. Meltwater from glacial calving could suppress water-column convection and decrease deep-sea oxygen; overturn⁶ could contribute to the low $\delta^{13}\text{C}$ in post-glacial sequences.

The idea that a modest CO_2 increase would eliminate tropical glaciation in the open-water solution assumes a linear response, which is in contrast to many examples in the Pleistocene and the model used in our study^{1,7}. Increases in CO_2 produce virtually no meltback until a threshold level is reached¹, whereupon a small increase causes rapid melting. Calculated ice-sheet retreat times of tens of thousands of years agree with sedimentological estimates⁸. Isostatic depression from the large ice sheets we describe could accommodate a thick layer of post-glacial carbonate (only the thin bottom layer of which is cap rock in the strictest sense).

We are not as sanguine as Schrag and Hoffman about the ability of metazoans to survive under the extreme conditions of a hard snowball Earth. Whether life could survive on a few scattered volcanic islands is a matter of conjecture. A "thin-ice" scenario⁹ is not consistent with results¹⁰ indicating that such regions have temperatures substantially colder than those estimated in ref. 9 (with implications for sea-ice thickness). Although evidence for life extends almost to the oldest rocks, three billion years transpired before the appearance of metazoans. This vast time interval suggests that the environmental tolerance of metazoans is much narrower than that of their simpler and hardier colleagues. If deep waters were anoxic, they could not survive on deep-sea chemosynthetic communities either, as these organisms still require free oxygen.

Future data may call for the reassessment of our open-water scenario, but we consider that the hard-snowball scenario is not yet proven. We believe that the open-water solution is much more favourable for the survival of metazoans, allowing their remote progeny to continue this discussion.

**William T. Hyde, Thomas J. Crowley,
Steven K. Baum, W. Richard Peltier***

Department of Oceanography, Texas A&M
University, College Station, Texas 77843, USA
e-mail: hyde@rossby.tamu.edu

*Department of Physics, University of Toronto,
Toronto, Ontario M5S 1A7, Canada

- Crowley, T. J., Hyde, W. T. & Peltier, W. R. *Geophys. Res. Lett.* **28**, 283–286 (2001).
- Sohl, L. E. *et al. Geol. Soc. Am. Bull.* **111**, 1120–1139 (1999).
- Jacobsen, S. B. & Kauffman, A. J. *Chem. Geol.* **161**, 37–57 (1999).
- Williams, G. E. *Sedim. Geol.* **106**, 165–175 (1996).
- Kennedy, M. J. *et al. Geology* **26**, 1059–1063 (1998).
- Knoll, A. *et al. Science* **273**, 452–457 (1996).
- Hyde, W. T. *et al. Clim. Dynam.* **15**, 619–629 (1999).
- Kennedy, M. J. & Christie-Blick, N. in *Soc. Sedim. Geol. Conf. on Strata and Sequences on Shelves and Slopes* (1998).
- McKay, C. P. *Geophys. Res. Lett.* **27**, 2153–2156 (2000).
- Baum, S. K. & Crowley, T. J. *Geophys. Res. Lett.* (in the press).

The hormone resistin links obesity to diabetes

Claire M. Steppan, Shannon T. Bailey, Savitha Bhat, Elizabeth J. Brown, Ronadip R. Banerjee, Christopher M. Wright, Hiralben R. Patel, Rexford S. Ahima & Mitchell A. Lazar

Division of Endocrinology, Diabetes, and Metabolism, Departments of Medicine and Genetics, and The Penn Diabetes Center, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

Diabetes mellitus is a chronic disease that leads to complications including heart disease, stroke, kidney failure, blindness and nerve damage. Type 2 diabetes, characterized by target-tissue resistance to insulin, is epidemic in industrialized societies and is strongly associated with obesity; however, the mechanism by which increased adiposity causes insulin resistance is unclear. Here we show that adipocytes secrete a unique signalling molecule, which we have named resistin (for resistance to insulin). Circulating resistin levels are decreased by the anti-diabetic drug rosiglitazone, and increased in diet-induced and genetic forms of obesity. Administration of anti-resistin antibody improves blood sugar and insulin action in mice with diet-induced obesity. Moreover, treatment of normal mice with recombinant resistin impairs glucose tolerance and insulin action. Insulin-stimulated glucose uptake by adipocytes is enhanced by neutralization of resistin and is reduced by resistin treatment. Resistin is thus a hormone that potentially links obesity to diabetes.

Diabetes mellitus is a principal cause of morbidity and mortality in human populations. A minority of patients suffer from type 1 diabetes mellitus, which is caused by pancreatic β -cell failure and leads to an absolute loss of insulin¹. Type 2 diabetes mellitus is far more common, and is swiftly increasing in prevalence in industrialized societies. Type 2 diabetes mellitus is characterized by target-tissue resistance to insulin that cannot be overcome by β -cell hypersecretion². Although type 2 diabetes mellitus is strongly associated with insulin resistance in both human and rodent models of type 2 diabetes mellitus³, the connection between increased adiposity and insulin resistance remains unknown⁴. Fat-cell secretion of free fatty acids may contribute to insulin resistance in peripheral tissues⁵. Adipocytes also secrete a variety of polypeptides, such as leptin, tumour necrosis factor- α (TNF- α), adipon and Acrp30/adipoQ, that may affect insulin action in other tissues^{6–11}.

A new class of anti-diabetic drugs called thiazolidinediones (TZDs) enhance target-tissue sensitivity to insulin *in vivo*¹² where they function as high-affinity ligands for a nuclear receptor, particularly abundant in fat cells, called peroxisome proliferator activated receptor- γ (PPAR γ)^{13–15}. The best understood function of PPAR γ is as an adipogenic determination factor^{16–18}. However, it is probable that PPAR γ is the biological target of the anti-diabetic actions of TZDs: there is a strong correlation between PPAR γ binding *in vitro* and glucose lowering *in vivo*¹⁹; non-TZD PPAR γ ligands also increase insulin sensitivity¹⁹; activators of the PPAR γ heterodimer partner, retinoid X receptor, also have anti-diabetic properties²⁰; and patients heterozygous for a dominant-negative PPAR γ allele have severe insulin resistance, providing a genetic link between PPAR γ and diabetes²¹. Notably, mice heterozygous for a null PPAR γ allele develop adipocyte hypertrophy on a high-fat diet and have increased sensitivity to insulin^{22,23}. The gene targets of TZD-bound PPAR γ that regulate insulin sensitivity are unknown.

We thought that insulin resistance might be mediated by a TZD-regulated adipocyte-derived factor. A screen for genes that are induced during adipocyte differentiation but downregulated in mature adipocytes exposed to TZDs led to the discovery of a protein that we have called resistin (Fig. 1a). Resistin gene expression is induced during adipocyte differentiation, and the resistin polypeptide is specifically expressed and secreted by adipocytes. Resistin circulates in mouse serum, and its level is increased markedly in both genetic and diet-induced obesity. Immunoneutralization

improves blood glucose and insulin action in this model of type 2 diabetes. By contrast, administration of resistin impairs glucose tolerance and insulin action in normal mice. Resistin is therefore a strong candidate to explain the anti-diabetic effects of TZDs, as well as a mechanism by which excess adiposity leads to insulin resistance.

Resistin is a TZD-regulated protein

We considered that TZD treatment downregulates an adipocyte gene that contributes to insulin resistance. Resistin messenger RNA is induced markedly during adipocyte differentiation of 3T3-L1 cells, similar to other adipocyte-specific genes, such as PPAR γ (Fig. 1b). Notably, resistin gene expression was markedly downregulated by treatment with the TZD rosiglitazone, unlike the adipocyte-specific protein aP2, a well-characterized PPAR γ target gene¹⁴ (Fig. 1c). Other TZDs, including pioglitazone and troglitazone, also downregulated resistin (data not shown).

A notable feature of the deduced amino-acid sequence of resistin is a hydrophobic amino terminus that is predicted by the Prosort³⁴ and SignalP³⁵ algorithms to be a functional export signal (Fig. 1a). Resistin was abundant in the media of transfected 293T cells, and N-terminal Edman sequencing showed that the protein begins at Ser 21 of the deduced amino-acid sequence (data not shown). This suggests that resistin is specifically processed and secreted. Of note, endogenous resistin was secreted into the media by 3T3-L1 adipocytes (Fig. 1d). Moreover, consistent with its gene expression, resistin secretion was reduced markedly after exposure of the adipocytes to rosiglitazone for 4 d (Fig. 1e).

Adipocyte-specific expression of resistin

We next evaluated resistin gene expression in the mouse. A single mRNA of roughly 750 residues was robustly expressed in white adipose tissue but not in several other mouse tissues (Fig. 2a and b). Resistin expression was greater in white adipose tissue than in brown adipose tissue, where resistin mRNA was barely detectable (Fig. 2b). A low level of resistin expression was detected in mammary tissue, which is potentially due to the presence of the mammary fat pad (Fig. 2a). Resistin mRNA levels varied as a function of white adipose depot and gender, with the highest level of expression in female gonadal fat (Fig. 2b). Immunohistochemistry of epididymal white adipose tissue showed that the resistin protein was abundant in adipocyte cytoplasm (Fig. 2c). We have

also identified a human expressed sequence tag that is highly related to resistin. The deduced amino-acid sequence includes an N-terminal signal sequence and strict conservation of cysteine residues (Fig. 2d). The unique pattern of cysteines (X₁₁-C-X₈-C-X-C-X₃-C-X₁₀-C-X-C-X-C-X₉-CC-X₃₋₆-END) is conserved in a family of resistin-like molecules including at least three distinct mouse subtypes^{24,25}. The human resistin homologue is also secreted after transfection into 293T cells (data not shown). We showed adipocyte expression of the putative human resistin by using polymerase chain reaction with reverse transcription (RT-PCR) analysis (data not shown). The coding region of the entire human resistin gene was localized to a cloned fragment of human chromosome 19 (accession number AC008763, data not shown).

Downregulation of resistin

We analysed serum resistin levels, as resistin is a secreted protein produced by adipocytes. Notably, immunoreactive resistin was readily detectable in serum from normal mice (Fig. 3a). Resistin was also detectable in normal rat serum (data not shown). Mouse serum levels decreased after a 48-h fast, and were reversed by refeeding (Fig. 3a). Resistin gene and protein expression were both downregulated in epididymal fat (Fig. 3b, c). Moreover, consistent with the *in vitro* results, rosiglitazone treatment markedly lowered serum resistin levels in mice (Fig. 3d).

Diet-induced and genetically obese mice

Resistin was identified in a search for an adipocyte signal that mediated target tissue insulin resistance. We therefore analysed serum resistin levels in mice that developed obesity and insulin resistance on a high-fat diet²⁶. Resistin serum levels were markedly elevated in mice fed on a diet containing 45% (by kJ) of fat for eight weeks (Fig. 4a). The increase in serum resistin was observed within four weeks of high-fat feeding, as mice became insulin resistant and obese (Fig. 4b). Serum resistin levels were also elevated in *ob/ob* and *db/db* mice in which obesity and diabetes are inherited traits (Fig. 4c).

Neutralization of resistin

We next considered whether the increased level of resistin in mice

with diet-induced obesity was causally related to the associated type 2 diabetes. As antibodies to other hormones and signalling molecules can neutralize their function, we tested the effects of administering anti-resistin immunoglobulin-γ (IgG) on blood glucose in mice with diet-induced obesity, insulin resistance and hyperglycaemia (mean glucose 1.63 ± 0.10 mg ml⁻¹). Treatment of mice with nonspecific IgG had no significant effect on blood glucose (Fig. 5a). By contrast, treatment of mice with an equivalent amount of IgG purified from anti-resistin antiserum caused a significant decrease in blood glucose (1.37 ± 0.98 mg ml⁻¹, $P = 0.02$). After IgG therapy, blood glucose returned to levels indistinguishable from those before the experiment (Fig. 5b). Neutralization of resistin reversibly reduced hyperglycaemia in this model of diet-induced insulin resistance. To investigate the mechanism governing this effect, insulin tolerance tests were performed after treatment with control or anti-resistin IgG. Mice treated with anti-resistin IgG showed significantly improved sensitivity to insulin compared with control IgG treatment (Fig. 5b). These data indicate that resistin may be involved in mediating insulin resistance associated with diet-induced obesity.

Resistin impairs glucose tolerance *in vivo*

To investigate resistin function *in vivo*, recombinant resistin was expressed with a carboxy-terminal Flag-tag (resistin-F) in 293T cells, and affinity purified from the medium to apparent homogeneity (Fig. 6a). This preparation, containing less than 1 endotoxin unit per ml, was used for *in vivo* studies. Serum resistin concentration increased 15 min after intraperitoneal (i.p.) administration of 16.5 μg purified, recombinant resistin-F (Fig. 6b). Resistin levels were maximally elevated at 30–60 min, and reduced but still elevated 4 h after i.p. administration (Fig. 6b).

We next measured glucose tolerance after i.p. administration of resistin-F. Purified resistin-F (16.5 μg i.p.) or vehicle was administered to C57Bl/6J mice at the beginning of an overnight fast and again 12 h later. We measured glucose tolerance 2 h after the second injection of resistin-F. Peak blood glucose increased by 28% ($P = 0.004$) in the resistin-treated mice (Fig. 6c). Insulin levels were increased concomitantly with the increased blood glucose in resistin-F-treated mice (data not shown). Although this

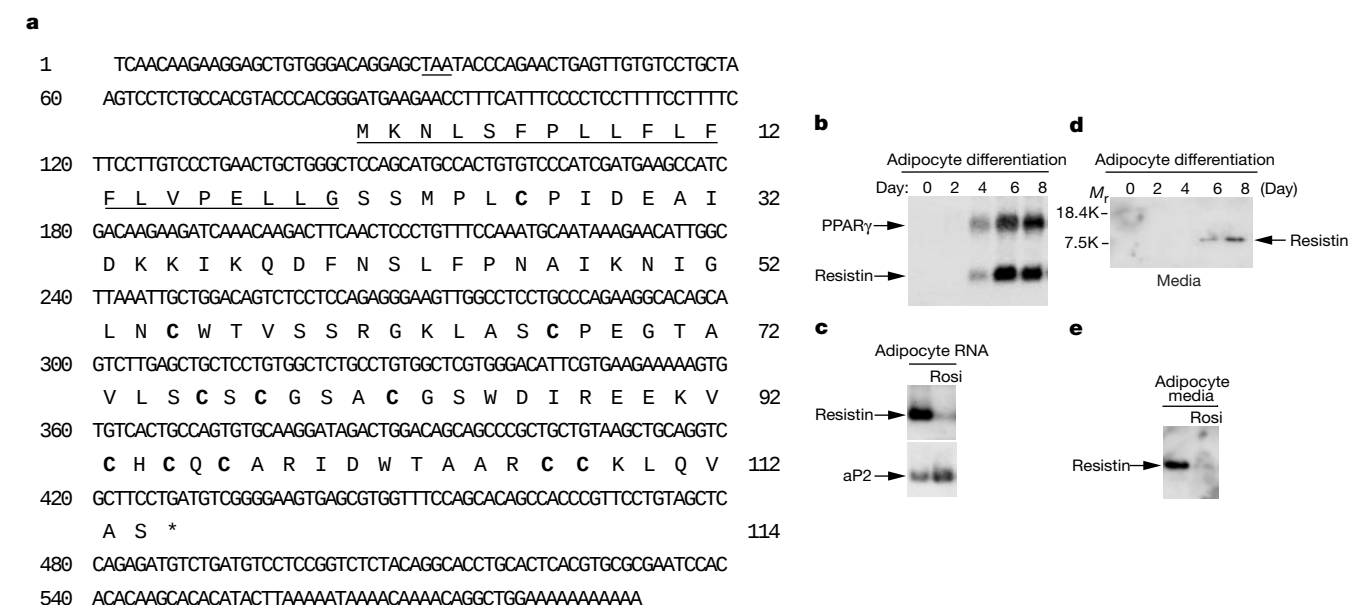


Figure 1 Cloning and identification of resistin. **a**, Resistin cDNA and amino-acid sequence. Stop codon, polyadenylation sequences and putative signal sequence are underlined. Asterisk indicates stop codon; cysteine residues are shown in bold. **b**, Northern analysis of resistin and PPARγ gene expression during adipogenic

differentiation of 3T3-L1 cells. **c**, Downregulation of resistin gene expression in 3T3-L1 adipocytes by rosiglitazone (Rosi, 1 μM). Expression of aP2 is shown for comparison. **d**, Resistin protein is induced and secreted during adipogenesis. **e**, Adipocyte resistin secretion is markedly reduced by rosiglitazone (Rosi, 1 μM).

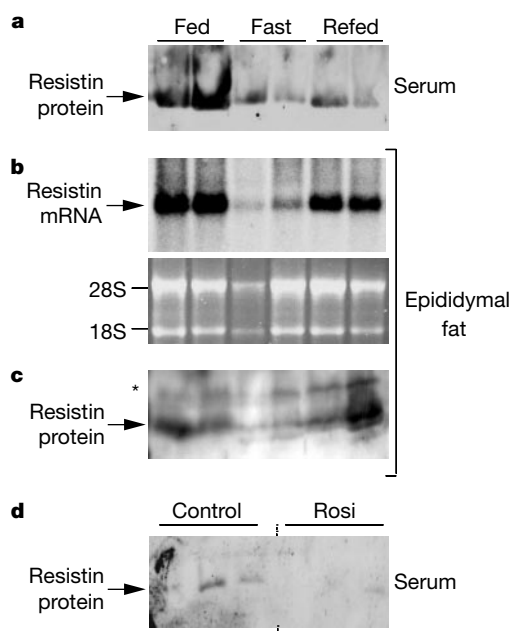


Figure 3 Resistin circulates in blood and is regulated by feeding and rosiglitazone. **a**, Resistin immunoblot of serum from C57Bl/6J mice with unrestricted feeding, fasted, or fasted and then refed. **b**, **c**, Adipocyte resistin gene expression (**b**) and protein (**c**) are reduced by fasting. Ribosomal RNA staining is shown as a control for loading of northern blot, and a nonspecific background band (asterisk) is nearly equal in all lanes of the immunoblot. **d**, Rosiglitazone treatment downregulates serum resistin levels.

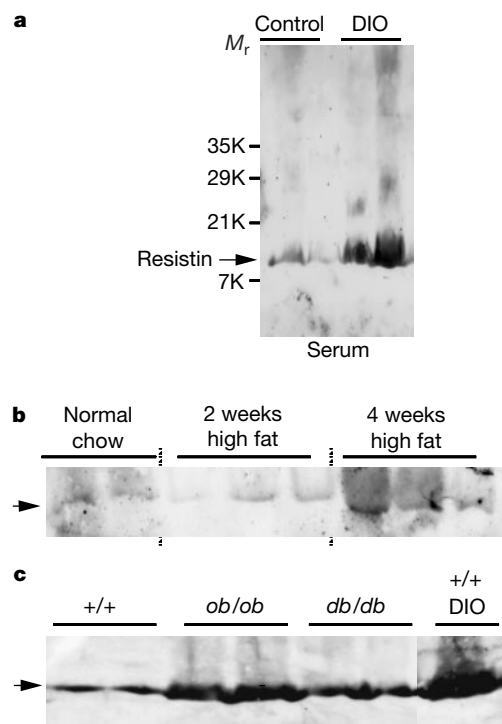


Figure 4 Resistin levels are elevated in obesity. **a**, Male AKR/J mice were fed a high-fat diet (DIO) or control chow from 4 to 12 weeks of age. Weight increased by 38% on the high-fat diet; serum insulin increased (2.1 ± 0.5 versus 5.8 ± 0.8 ng ml⁻¹). **b**, Time course. Male C57Bl/6J mice were fed normal chow (glucose 0.94 ± 0.04 mg ml⁻¹, insulin 0.47 ± 0.06 ng ml⁻¹) or a high-fat diet for 2 weeks (glucose 1.21 ± 0.07 mg ml⁻¹, insulin 2.6 ± 0.58 ng ml⁻¹) and 4 weeks (glucose 1.36 ± 0.04 mg ml⁻¹, insulin 3.6 ± 0.5 ng ml⁻¹). **c**, Resistin level in 10 μ l of serum from C57Bl/6J wild-type, *ob/ob*, *db/db* and diet-induced obesity (DIO) mice.

expression *in vivo* is specific to white adipose tissue, and resistin is found in the serum of normal mice. Resistin levels are increased in diet-induced obesity as well as in genetic models of obesity and insulin resistance. Resistin is therefore a candidate adipocyte-derived factor that contributes to insulin resistance *in vivo*. This hypothesis is supported by studies in adipocytes, where neutralization with resistin antiserum enhanced insulin-stimulated glucose uptake, and insulin action was blunted by recombinant resistin. Administration of resistin to mice impaired glucose tolerance without reducing insulin levels, and decreased sensitivity to the effects of insulin. Neutralization of resistin reduced the hyperglycaemia of obese, insulin-resistant mice, at least in part, by improving sensitivity to insulin. These data suggest that resistin is a unique hormone whose effects on glucose metabolism are antagonistic to those of insulin.

The effects of resistin administration and neutralization on glucose tolerance *in vivo* indicate that muscle, in addition to fat, may be a target for resistin; however, the molecular target of resistin is unknown. We speculate that resistin signals, by means of a receptor in insulin-responsive tissues, to modulate one or more steps in the insulin signalling pathway. Resistin may also have effects on tissues other than fat and muscle, such as liver and brain, and could have several functions. As with leptin²⁷, it seems probable that resistin did not evolve specifically to cause insulin resistance during times when food is plentiful. Rather, its physiological function may also be in the adaptive response to starvation. Additional work is required to understand better the biological function of resistin.

PPAR γ ligands represent an important breakthrough in the therapy of type 2 diabetes, and their ability to downregulate resistin suggests that resistin may be a link between obesity, diabetes and the

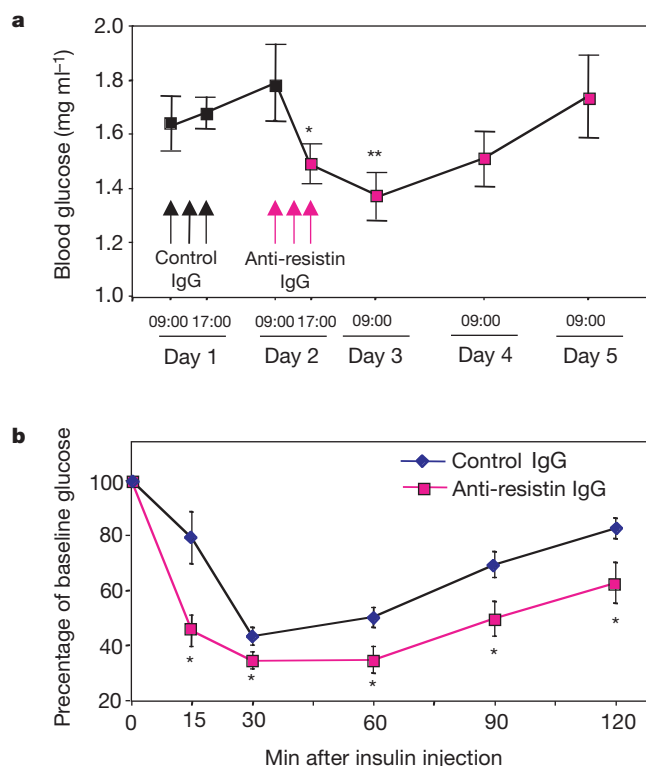


Figure 5 Neutralization of resistin improves hyperglycaemia and insulin resistance in mice. **a**, Glucose levels. C57Bl/6J mice were fed a high-fat diet from 8 to 16 weeks of age, then sequentially given three i.p. injections of 500 μ g control rabbit IgG or anti-resistin IgG (arrows). Data are mean $\pm$ s.e.m. ($n = 4$). Asterisk, $P = 0.04$; double asterisk, $P = 0.02$ (paired *t*-test). **b**, Insulin tolerance testing. Mice received three i.p. injections of 1,000 μ g control rabbit IgG or anti-resistin IgG (see Methods). Data are mean $\pm$ s.e.m. ($n = 6$) of the percentage of baseline glucose. Asterisk, $P = 0.0001$ (paired *t*-test).

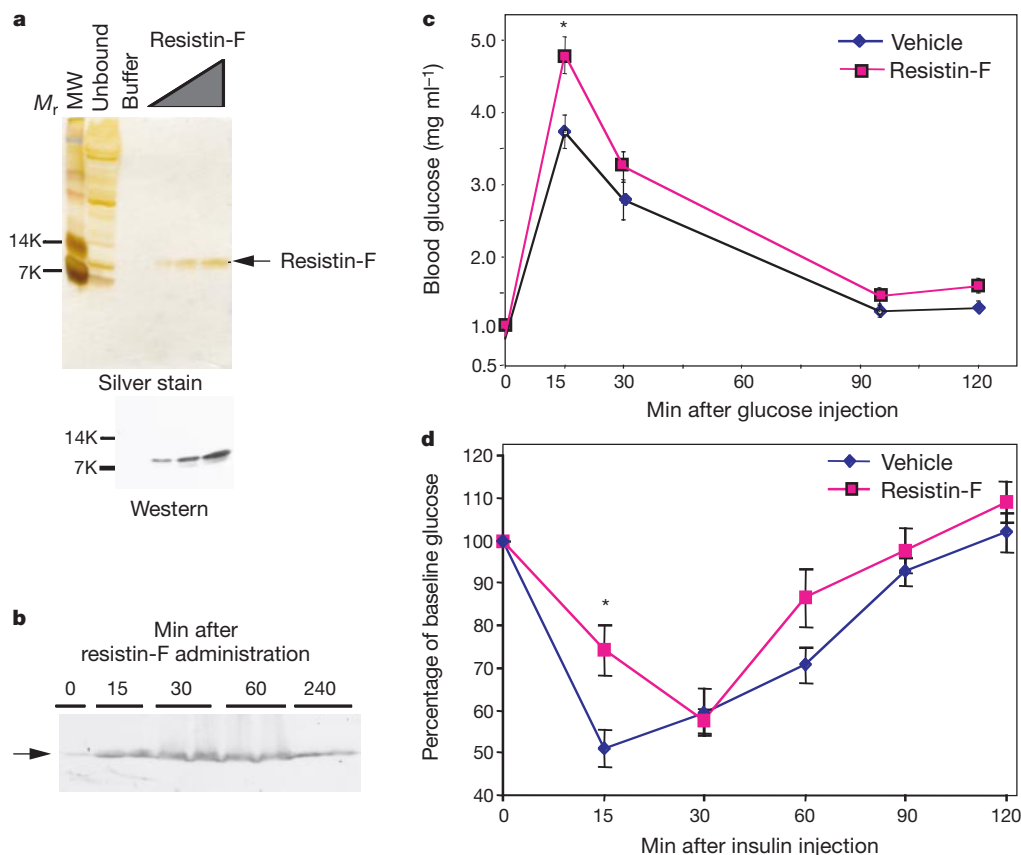


Figure 6 Resistin impairs glucose tolerance and insulin action in mice. **a**, Silver stain (top) and resistin immunoblot of 1, 2, and 4 μg purified resistin-F. MW, molecular mass markers. **b**, Serum levels of resistin at indicated times after a single dose of resistin-F (16.5 μg i.p.). **c**, Glucose tolerance. Resistin-F (16.5 μg) was administered i.p. to C57BL/

6J mice 14 h and 2 h before glucose administration. Data are mean $\pm$ s.e.m. ($n = 15$ per group). Asterisk, $P = 0.004$ (unpaired t -test). **d**, Insulin tolerance. Mice were treated as in **c**, except that four doses of resistin-F were given. Data are mean $\pm$ s.e.m. ($n = 9$ –10 per group) of the percentage of baseline glucose. Asterisk, $P = 0.003$ (unpaired t -test).

mechanism of action of these anti-diabetic drugs. Potential complications of PPAR γ activation in tissues other than fat^{28,29} may conceivably be avoided by making resistin the target of anti-diabetic therapy if the regulation and properties of human resistin are similar to those of mouse resistin. Such therapies could include

reduction of serum resistin level, neutralization of the biological activity of circulating resistin, and/or antagonism of resistin action directed against the cellular receptor(s). □

Methods

Identification and cloning of resistin

We differentiated 3T3-L1 cells as described³⁰. Adipocytes were treated for 48 h with either dimethyl sulphoxide or rosiglitazone (1 μM in dimethyl sulphoxide). Complementary DNA was prepared from 1 μg of RNA using the Cap finder cDNA synthesis kit (Clontech Labs, Palo Alto, California). We used PCR-select (Clontech) to select for adipocyte genes expressed in the absence of rosiglitazone. The cDNAs were arrayed on 96-well plates and screened using labelled cDNA from adipocytes or from adipocytes treated with rosiglitazone. Northern analysis of 20 μg of total RNA was performed as described¹⁵.

Resistin antibody

Resistin was expressed in bacteria as a glutathione S-transferase fusion protein that was used to generate rabbit polyclonal antiserum.

Immunoblot analysis

We performed immunoblot assays on whole-cell extracts, medium or serum as described³¹, except that we applied samples to 15% SDS-PAGE in loading buffer containing 20% β -mercaptoethanol. We used rabbit resistin antiserum at a dilution of 1:500 or 1:1,000.

Immunohistochemistry

Epididymal fat from a CD1 mouse was fixed in 4% paraformaldehyde in PBS. We analysed paraffin-embedded sections using rabbit anti-resistin polyclonal antiserum (1:500) and goat antirabbit antibody (1:200). We counterstained slides with Gill's haematoxylin number 2 (Fisher Scientific, Fairlawn, New Jersey).

Expression and purification of resistin and resistin-F

Wild-type resistin and resistin-F (with C-terminal Flag epitope created using PCR) cDNAs were subcloned into expression vectors pCMX or pZac2.1 and transfected into 293T cells

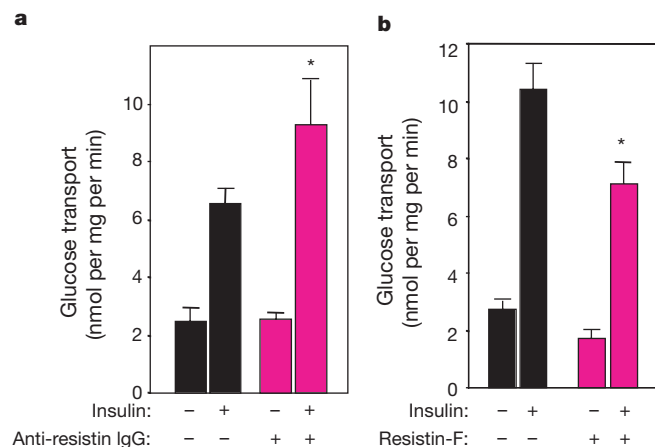


Figure 7 Resistin antagonizes insulin-stimulated glucose transport *in vitro*. **a**, 3T3-L1 adipocytes were stimulated with insulin (5 nM) after pretreatment with either control IgG or IgG purified from anti-resistin serum, and glucose uptake was measured. Mean $\pm$ s.e.m. ($n = 3$) are shown for each data point. Asterisk, $P < 0.04$. **b**, 3T3-L1 adipocytes were treated with resistin-F (1–2 μM) or vehicle, with or without insulin (10 nM), and glucose uptake was measured. Mean $\pm$ s.e.m. ($n = 6$) are shown for each data point. Asterisk, $P = 0.02$. The experiment shown is representative of three experiments.

using Fugene (Boehringer Mannheim). Resistin-F was batch-purified from conditioned media using Flag-Agarose (Sigma, St Louis, Missouri) and eluted with Flag peptide ($100 \mu\text{g ml}^{-1}$) in Tris-buffered saline. Pyrogen-free water was used in all steps of purification, and final endotoxin concentration of purified resistin-F and elution buffer control were determined to be less than 1 endotoxin unit per ml using the limulus amoebocyte lysate assay. We performed N-terminal sequencing of purified, secreted resistin-F by Edman reaction at the Wistar Protein Microchemistry Facility³².

In vivo experiments

Animal care and procedures were in accordance with guidelines and regulations of the institutional animal care and use committee of the University of Pennsylvania.

For fasting and rosiglitazone treatment, 2 weeks after acclimatization, 8-week-old male C57Bl/6J mice (Jackson Labs, Bar Harbor, ME) were housed ($n = 4$ per cage) under 12-h light/dark cycles (lights on at 06:00), ambient temperature 23°C , and allowed unrestricted access to chow and water. One group was maintained on chow, a second was fasted for 48 h, and a third was fasted for 48 h followed by unrestricted access to chow 48 h. We dissected epididymal fat for RNA and protein analysis. For rosiglitazone treatment experiments, female FVB/N mice, initially 5 weeks old, were treated with rosiglitazone ($0.012 \text{ mg per g diet}$; $\sim 3 \text{ mg per kg per day}$) for five weeks. Sera from these mice were provided by L. Chao and M. Reitman (Diabetes Branch, National Institutes of Diabetes, Digestive, and Kidney Diseases (NIDDK), Bethesda, Maryland).

For antibody neutralization, the effect of anti-resistin IgG on blood glucose was analysed in male C57Bl/6J mice fed on a high-fat diet from 8 to 16 weeks of age. After induction of diet-induced obesity, mice received three i.p. injections of sterile rabbit IgG ($500 \mu\text{g}$ in $50 \mu\text{l}$ 0.02 M potassium phosphate, 150 mM NaCl, $\text{pH } 7.2$) at 09:00, 12:00 and 15:00. The next day they received three i.p. injections of anti-resistin IgG ($500 \mu\text{g}$ in $50 \mu\text{l}$) at the above times. Tail bleeds were performed for glucose measurement at 09:00 and 17:00 on the day of treatment, and then daily for 3 d thereafter. The same protocol was used to evaluate the effect of anti-resistin IgG on insulin tolerance except that the dose of IgG was 1 mg . After the third dose of IgG, mice were fasted for 4 h before i.p. administration of insulin (0.75 U per kg Novolin R (Novo Nordisk, Princeton, New Jersey)).

To test the effects of resistin-F, serum resistin levels were analysed in normal chow-fed littermates at 0, 15, 30, 60 and 240 min after i.p. injection of resistin-F ($16.5 \mu\text{g}$ in Flag elution buffer). For glucose tolerance tests, male and female C57Bl/6J mice aged 5–8 weeks received resistin-F ($16.5 \mu\text{g}$) or vehicle i.p. before starting an overnight fast, and a second dose 12 h later. Two hours after the second dose, glucose (2 g kg^{-1} i.p.) was administered and glucose levels were measured in tail blood at indicated times. For insulin tolerance tests, mice received four doses of resistin-F or vehicle at 08:00, 14:00, 20:00 and 08:00 the next day, before a 4-h fast and administration of insulin (0.5 U per kg) and measurement of blood glucose.

Glucose transport assay

Glucose transport was assayed by measuring [^3H]2-deoxyglucose uptake as described³³. Twelve hours after serum starvation, cells were treated with insulin for 15 min at the indicated concentrations. Resistin rabbit polyclonal antibody (IgG) or IgG control was added at $9.4 \mu\text{g ml}^{-1}$ at the time of serum starvation. Resistin-F ($1\text{--}2 \mu\text{M}$) or Flag elution buffer was added at the time of serum starvation.

Received 9 June; accepted 15 November 2000.

- Nathan, D. M. Long-term complications of diabetes mellitus. *New Eng. J. Med.* **328**, 1676–1685 (1993).
- Taylor, S. I. Deconstructing type 2 diabetes. *Cell* **97**, 9–12 (1999).
- Kopelman, P. G. Obesity as a medical problem. *Nature* **404**, 635–643 (2000).
- Kahn, C. R., Vicent, D. & Doria, A. Genetics of non-insulin-dependent (type II) diabetes mellitus. *Annu. Rev. Med.* **47**, 509–531 (1996).
- Boden, G. Role of fatty acids in the pathogenesis of insulin resistance and NIDDM. *Diabetes* **46**, 1–10 (1997).
- Hotamisligil, G. S. The role of TNF α and TNF receptors in obesity and insulin resistance. *J. Int. Med.* **245**, 621–625 (1999).
- Spiegelman, B. M. & Flier, J. S. Adipogenesis and obesity: rounding out the big picture. *Cell* **87**, 377–389 (1996).
- Mohamed-Ali, V., Pinkney, J. H. & Coppack, S. W. Adipose tissue as an endocrine and paracrine organ. *Int. J. Obes. Relat. Metab. Disord.* **22**, 1145–1158 (1998).
- Friedman, J. M. & Halaas, J. L. Leptin and the regulation of body weight in mammals. *Nature* **395**, 763–770 (1998).
- Shimomura, I., Hammer, R. E., Ikemoto, S., Brown, M. S. & Goldstein, J. L. Leptin reverses insulin

- resistance and diabetes mellitus in mice with congenital lipodystrophy. *Nature* **401**, 73–76 (1999).
- Moller, D. E. Potential role of TNF α in the pathogenesis of insulin resistance and type 2 diabetes. *Trends Endocrinol. Metab.* **11**, 212–217 (2000).
- Henry, R. R. Thiazolidinediones. *Endocrinol. Metab. Clin. North Am.* **26**, 553–573 (1997).
- Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for the nuclear peroxisome proliferator-activated receptor γ (PPAR γ). *J. Biol. Chem.* **270**, 12953–12956 (1995).
- Tontonoz, P., Hu, E., Graves, R. A., Budavari, A. I. & Spiegelman, B. M. mPPAR γ 2: tissue-specific regulator of an adipocyte enhancer. *Genes Dev.* **8**, 1224–1234 (1994).
- Chawla, A., Schwarz, E. J., Dimaculangan, D. D. & Lazar, M. A. Peroxisome proliferator-activated receptor γ (PPAR γ): Adipose predominant expression and induction early in adipocyte differentiation. *Endocrinology* **135**, 798–800 (1994).
- Barak, Y. *et al.* PPAR γ is required for placental, cardiac, and adipose tissue development. *Mol. Cell* **4**, 585–595 (1999).
- Rosen, E. D. *et al.* PPAR γ is required for the differentiation of adipose tissue *in vivo* and *in vitro*. *Mol. Cell* **4**, 611–617 (1999).
- Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPAR γ 2, a lipid-activated transcription factor. *Cell* **79**, 1147–1156 (1994).
- Willson, T. M., Brown, P. J., Sternbach, D. D. & Henke, B. R. The PPARs: from orphan receptors to drug discovery. *J. Med. Chem.* **2000**, 527–550 (2000).
- Mukherjee, R. *et al.* Sensitization of diabetic and obese mice to insulin by retinoid X receptor agonists. *Nature* **386**, 407–410 (1997).
- Barroso, I. *et al.* Dominant negative mutations in human PPAR γ associated with severe insulin resistance, diabetes mellitus, and hypertension. *Nature* **402**, 880–883 (1999).
- Kubota, N. *et al.* PPAR γ mediates high-fat diet-induced adipocyte hypertrophy and insulin resistance. *Mol. Cell* **4**, 597–609 (1999).
- Miles, P. D., Barak, Y., He, W., Evans, R. M. & Olefsky, J. M. Improved insulin-sensitivity in mice heterozygous for PPAR- γ deficiency. *J. Clin. Invest.* **105**, 287–292 (2000).
- Holcomb, I. N. *et al.* FIZZ1, a novel cysteine-rich secreted protein associated with pulmonary inflammation, defines a new gene family. *EMBO J.* **19**, 4046–4055 (2000).
- Steppan, C. M. *et al.* A family of tissue-specific resistin-like molecules. *Proc. Natl Acad. Sci. USA* (in the press).
- VanHeek, M. *et al.* Diet-induced obese mice develop peripheral, but not central, resistance to leptin. *J. Clin. Invest.* **99**, 385–390 (1997).
- Ahima, R. S. *et al.* Role of leptin in the neuroendocrine response to fasting. *Nature* **382**, 250–252 (1996).
- Seed, B. PPAR γ and colorectal carcinoma: conflicts in a nuclear family. *Nature Med.* **4**, 1004–1005 (1998).
- Tontonoz, P., Nagy, L., Alvarez, J. G., Thomazy, V. A. & Evans, R. M. PPAR γ promotes monocyte/macrophage differentiation and uptake of oxidized LDL. *Cell* **93**, 241–252 (1998).
- Shao, D. & Lazar, M. A. PPAR γ , C/EBP α , cell cycle status and the commitment to adipocyte differentiation. *J. Biol. Chem.* **272**, 21473–21478 (1997).
- Huang, E. Y. *et al.* Nuclear receptor corepressors partner with class II histone deacetylases in a Sin3-independent repression pathway. *Genes Dev.* **14**, 45–54 (2000).
- Speicher, D. W. & Reim, D. in *Current Protocols in Protein Science* (eds Coligan, J. E., Dunn, B. M., Ploegh, H. L., Speicher, D. W. & Wingfield, P. T.) 11.10.11–11.10.38 (John Wiley & Sons, New York, 1997).
- Hausdorff, S. F. *et al.* Identification of wortmannin-sensitive targets in 3T3-L1 adipocytes. *J. Biol. Chem.* **274**, 24677–24684 (1999).
- Nakai, K. & Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34–36 (1999).
- Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engng* **10**, 1–6 (1997).

Acknowledgements

We thank D. Shao for help with the early stages of this project, and M. Brown, M. S. Brown, J. Cunningham, T. Lawrence, M. Birnbaum, J. Stephens, A. Swick and the Lazar laboratory for discussions. We acknowledge D. Reim and D. Speicher of the Wistar Protein Microchemistry/MS Facility for sequence analysis; H. Collins and the Radioimmunoassay Core of the Penn Diabetes Center and J. Moffett; and G. Swain, and the Morphology Core of the Penn Center for the Molecular Study of Digestive Diseases. This work was supported by NIDDK grants to M.A.L., and by the Penn Diabetes Center. C.M.S. was supported by an unrestricted postdoctoral research fellowship from Pfizer. E.J.B. was supported by a student research fellowship from the American Diabetes Association. R.R.B. is a trainee of the Medical Scientist Training Program.

Correspondence and requests for materials should be addressed to M.A.L. (e-mail: lazar@mail.med.upenn.edu). The mouse and human resistin proteins are deposited under GenBank accession numbers AF323080 and AF323081, respectively.

Non-detection at Venus of high-frequency radio signals characteristic of terrestrial lightning

D. A. Gurnett*, P. Zarka†, R. Manning†, W. S. Kurth*, G. B. Hospodarsky*, T. F. Averkamp*, M. L. Kaiser‡ & W. M. Farrell‡

* Department of Physics and Astronomy, The University of Iowa, Iowa City, Iowa 52242, USA

† Space Research Department, Observatoire de Paris, CNRS UMR 8632, Meudon, France

‡ NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

The detection^{1,2} of impulsive low-frequency (10 to 80 kHz) radio signals, and separate very-low-frequency (~100 Hz) radio 'whistler' signals³⁻⁵ provided the first evidence for lightning in the atmosphere of Venus. Later, a small number of impulsive high-frequency (100 kHz to 5.6 MHz) radio signals, possibly due to lightning, were also detected⁶. The existence of lightning at Venus has, however, remained controversial⁷⁻¹³. Here we report the results of a search for high-frequency (0.125 to 16 MHz) radio signals during two close fly-bys of Venus by the Cassini spacecraft. Such signals are characteristic of terrestrial lightning, and are commonly heard on AM (amplitude-modulated) radios during thunderstorms. Although the instrument easily detected signals from terrestrial lightning during a later fly-by of Earth (at a global flash rate estimated to be 70 s⁻¹, which is consistent with the rate expected for terrestrial lightning), no similar signals were detected from Venus. If lightning exists in the venusian atmosphere, it is either extremely rare, or very different from terrestrial lightning.

The Cassini spacecraft, which is on its way to Saturn, made two gravity-assisted fly-bys of Venus, the first (Venus-1) on 26 April 1998, and the second (Venus-2) on 24 June 1999. During these fly-bys the Radio and Plasma Wave Science instrument¹⁴ (RPWS) conducted a search for impulsive high-frequency radio signals from lightning. Impulsive signals of this type are commonly heard on AM radios during terrestrial thunderstorms, and are called spherics^{15,16}. Because spherics are always produced by terrestrial lightning, the detection of spherics by a spacecraft flying over another planet provides strong evidence of lightning. To be detected the signals must pass through the ionosphere. Such line-of-sight propagation can only occur at frequencies above the maximum plasma frequency in the ionosphere, which on the dayside of Venus is about 5 MHz, and on the nightside is about 1 MHz, or less¹⁷. The spectrum of lightning generally decreases with increasing frequency, approximately as 1/f² for terrestrial lightning; the best place to search for lightning is therefore on the nightside, where the ionospheric plasma frequency is low. Lightning can also be detected on the dayside, but the frequency must be above 5 MHz. To illustrate the region sampled by Cassini, the trajectories of the two Venus fly-bys are shown in Fig. 1. As can be seen the trajectories are quite complementary. The Venus-1 fly-by provided a low-altitude pass over the nightside of Venus, and the Venus-2 fly-by provided a low-altitude pass over the dayside of Venus. When the two fly-bys are combined, line-of-sight coverage is provided over 61% of the surface (in a Sun-fixed coordinate system) at radial distances less than 2 Venus radii, R_V, and 88% at radial distances less than 5 R_V. The coverage is reduced somewhat by refraction if the frequency is near the ionospheric plasma frequency.

The electric field intensities observed by the RPWS during the Venus-1 fly-by are shown in Fig. 2. The nearly constant background in all except the 0.125-kHz channel is due to galactic radio noise¹⁸.

About a dozen impulsive events can be seen near closest approach that could possibly be attributed to lightning. The strongest such event (with a peak intensity 4.1 dB above the galactic background) occurs in the 2.225-MHz channel at about 13:46 UT. Although at first glance this event looks like lightning, a detailed examination shows that it consists of a series of consecutive points in a single-frequency channel, with no corresponding response in the adjacent frequency channels (see the expanded panel in Fig. 2). But lightning produces radiation over a broad range of frequencies, so this event cannot be lightning. Several of the other events observed around closest approach have similar narrowband characteristics, and cannot be caused by lightning.

If the intensity scale in Fig. 2 is further magnified, a continuous level of fluctuations can be seen with a standard deviation of about $\sigma = 0.23$ dB. A detailed analysis of these fluctuations has been carried out by counting the number of peaks whose intensity, I , exceeds a threshold, I^* , that is a fixed number of standard deviations,

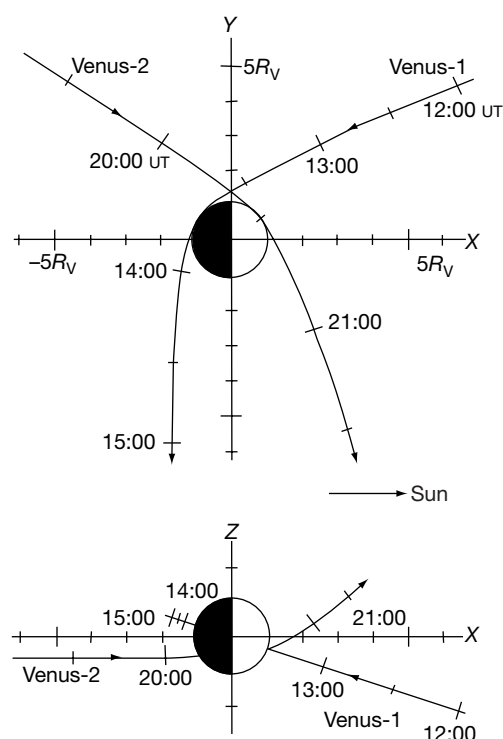


Figure 1 The trajectories of the Venus-1 and the Venus-2 fly-bys in a Venus-centred coordinate system. The +x-axis is directed toward the Sun, the +y is in the plane of Venus' orbit, and the +z axis completes the orthogonal coordinate system. The closest approach for Venus-1 occurred at an altitude of 284 km, and the closest approach for Venus-2 occurred at an altitude of 598 km. For the lightning search, measurements were made using a 18.5-m tip-to-tip electric dipole antenna. Intensities were measured in 20 quasi-logarithmically spaced frequency channels from 0.125 to 15.975 MHz, each with a bandwidth of 25 kHz. The receivers provided outputs proportional to the logarithm of the signal intensity, with response times of about 0.1 to 0.4 ms for rising signals and about 0.5 to 1.0 ms for falling signals. During the Venus-1 fly-by the receiver outputs were digitally averaged over 80-ms intervals, with 1 sample from each channel every 1.94 s. Data were collected from 12:58:17 Universal Time (UT) to 15:12:39 UT, giving a total of 81,920 measurements. During the Venus-2 fly-by the lowest 12 channels (0.125 to 3.975 MHz) were digitally averaged over 21-ms intervals and the upper 8 channels (5.475 to 15.975 MHz) were digitally averaged over 12-ms intervals, with 1 sample from each channel every 0.625 s. Data were collected from 19:15:57 UT to 21:19:49 UT, giving a total of 172,160 measurements. In addition, for the Venus-2 fly-by three frequency channels, 1.525, 5.325 and 9.125 MHz, were sampled at a rate of 1,000 samples per second (called the millisecond mode), with one block of 1,024 samples every 1.375 s. The millisecond-mode data were collected from 20:00:17 UT to 20:59:55 UT, giving a total of 1,307,648 measurements in the millisecond mode.

N_σ , above a locally averaged background, I_0 . The open squares in Fig. 3 shows the number of events with $I > I^*$ as a function of N_σ . The solid line is the expectation for a gaussian distribution. As can be seen, the intensities are in good agreement with simple random fluctuations in the galactic background. Only 12 peaks exceed four standard deviations, some of which are the previously mentioned narrowband signals. To search for very weak lightning signals we investigated the number of peaks above three standard deviations as a function of frequency and time for the entire data set. The resulting distribution shows no evidence of the $1/R^2$ radial dependence or the $1/f^2$ frequency dependence that would be expected for lightning. Also, there is no evidence of a local time variation in the occurrence rate, as one might expect from the day/night asymmetry in the ionospheric plasma frequency.

The statistical distribution of the intensity peaks for the Venus-2 fly-by is shown by the open circles in Fig. 3. Again the fluctuations closely follow a gaussian distribution, with no statistically significant peaks. In the high time-resolution millisecond-mode data (see the caption of Fig. 1 for a description of the millisecond mode), approximately 40 small widely scattered pulses were found with durations of about 1 to 2 milliseconds. Although these pulses are suggestive of lightning, they do not display the $1/R^2$ radial intensity variation and the $1/f^2$ frequency dependence that would be expected for lightning. Furthermore, a substantial number of the pulses occurred near closest approach on the dayside of Venus at a

frequency (1.525 MHz) that is well below the plasma frequency (~ 5 MHz) in the dayside ionosphere. These millisecond-mode pulses cannot be due to lightning, as the signals cannot propagate through the ionosphere. Most probably they are caused by interference from on-board electrical systems.

About two months after the Venus-2 fly-by, on 18 August 1999, the Cassini spacecraft made a close fly-by of Earth. To demonstrate that the RPWS can detect lightning, a search was conducted for impulsive signals from terrestrial lightning. More frequency channels (75) were used to aid in eliminating signals from terrestrial radio stations. Otherwise the analysis procedure was essentially the same as for the Venus-1 and Venus-2 data. The statistical distribution of intensity peaks is shown by the black dots in Fig. 3. As can be seen, many impulsive signals were detected above the gaussian background noise (1,101 events greater than three standard deviations). These impulsive signals were observed essentially continuously at all radial distances inside of about 14 Earth radii (R_E). The maximum rate occurred on the nightside of the Earth at about $4.5 R_E$, and was about 30 pulses per minute. After correcting for the instrument dead time and various other effects, such as obscuration by other types of radio signals, the global average flash rate is estimated to be about 70 s^{-1} , which is consistent with the global average flash rate expected for terrestrial lightning, that is about 100 s^{-1} . Some of the peaks were as much as 40 dB above the background noise level, and the intensities increased systematically with decreasing radial distance and frequency, varying approximately as $1/R^2$ and $1/f^2$, as would be expected for terrestrial lightning. Also, a strong day/night asymmetry is present in the low-frequency cut-off of the frequency spectrum, as would be expected from the day/night variation of the ionospheric plasma frequency.

The above comparisons between Venus and Earth show that if lightning exists on Venus it is either extremely rare, or very different from terrestrial lightning. Because instrumental characteristics play an important role in the ability to detect lightning, it is important to quantify the constraints implied by this study. Radio signals cannot propagate through the ionosphere at frequencies below the maximum plasma frequency in the ionosphere, which at its lowest is about 1 MHz on the nightside of Venus, so no definitive statement can be made about the lightning spectrum at frequencies below about 1 MHz. However, above 1 MHz, the constraints implied by this study are very restrictive, and are summarized by the diagram in

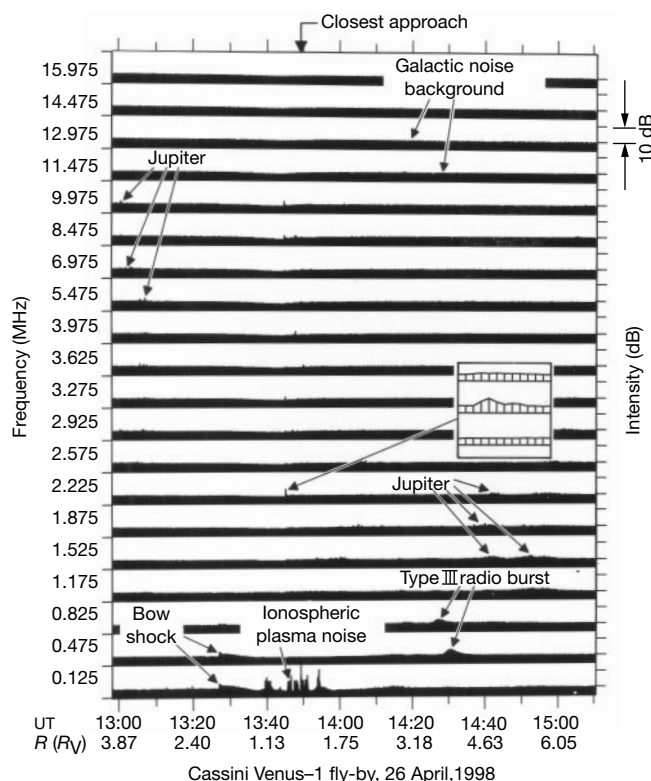


Figure 2 A 20-channel plot of the electric field intensities as a function of time during the Venus-1 fly-by. The frequencies are shown on the left, and the intensities in decibels are shown on the right. The UT and radial distance (R) from Venus, in Venus radii (R_V), are shown at the bottom of the plot. The background noise levels in all except the 0.125 kHz channel are due to galactic radio noise¹⁸, and have been adjusted to the same level in each channel (5 dB above the bottom of the plot). Features not associated with lightning that can be easily identified included some very weak emissions from Jupiter in the 5.475 to 9.975 MHz channels from about 13:00 to 13:10 UT, and again in the 1.175 to 2.225 MHz channels from about 14:35 to 15:00 UT (these can be seen in similar spectrograms from Earth-orbiting spacecraft), a type III solar radio burst in the 0.475 and 0.825 MHz channels from about 14:23 to 14:35 UT, and various types of ionospheric plasma wave noise in the 0.125 and 0.475 MHz channels from about 13:24 to 13:55 UT.

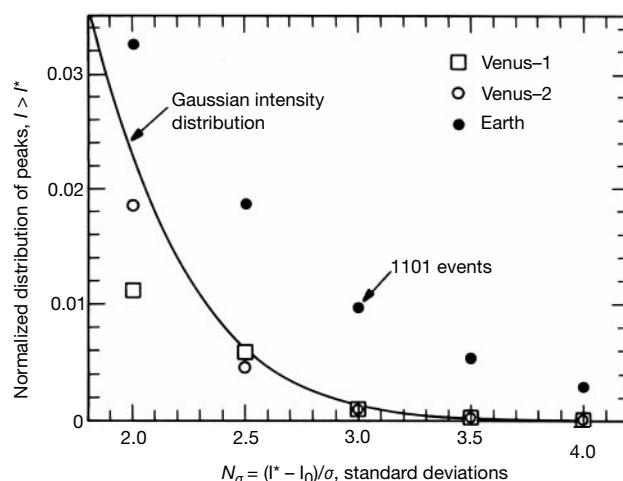


Figure 3 A plot of the number of intensity peaks greater than a fixed number of standard deviations above the background noise level, for the Venus-1, Venus-2, and Earth fly-bys. For the Venus-1 and Venus-2 fly-bys the number of peaks at each level is consistent with a gaussian distribution of fluctuations in the background noise. For the Earth fly-by a large number of impulsive events were detected over the entire region within about $13 R_E$, some with intensities as much as 40 dB above the background noise level. These impulses are consistent with the detection of terrestrial lightning.

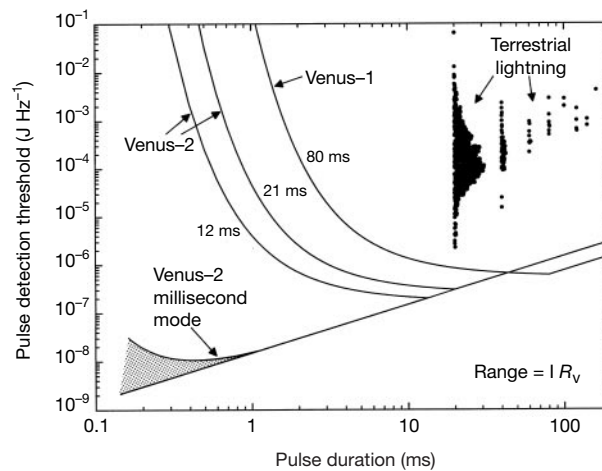


Figure 4 A plot of the pulse-detection threshold for the RPWS, in units of J Hz^{-1} , as a function of the pulse duration at a representative range of one Venus radii ($1 R_V$). The pulse detection threshold has been computed by multiplying the galactic noise level, in units of $\text{Watts m}^{-2} \text{Hz}^{-1}$, by $4\pi(R_V)^2$, by the pulse duration, and then by a factor of 0.236, which sets the pulse detection threshold at four standard deviations above the fluctuation level of the galactic noise background. The galactic noise level has been taken to be $7.5 \times 10^{-20} \text{ W m}^{-2} \text{Hz}^{-1}$, which is typical for the frequency range from about 1 to 10 MHz. Four curves are shown, one for Venus-1, with an integration time of 80 ms, and three for Venus-2, two with integration times of 12 ms (for the 0.125 to 3.975 channels) and 21 ms (for the 5.475 to 15.975 MHz channels) and one for the millisecond mode.

Fig. 4. This diagram shows the RPWS pulse-detection threshold, expressed in energy spectral density (in units of J Hz^{-1}), as a function of the duration of the pulse. Because the detection threshold depends on the distance to the source, a range of $1 R_V$ has been assumed. The threshold can be easily adjusted to other ranges using a simple $1/R^2$ law. Also shown for comparison are the energy spectral densities and pulse durations of the 1,101 lightning events detected during the Earth fly-by. As can be seen, the energy spectral density of terrestrial lightning extends as much as 40 dB above the energy detection thresholds at Venus for a range of $1 R_V$, and even more near closest approach. If terrestrial-like lightning were occurring in the atmosphere of Venus within the region viewed by Cassini, it would have been easily detectable.

Since the atmosphere of Venus is very different from that of Earth, it is perhaps not surprising that electrical activity on Venus might be very different from lightning in the Earth's atmosphere. Earth lightning can be divided into two broad types, cloud-to-ground, and cloud-to-cloud (including intra-cloud). Because clouds over Venus are at very high altitudes of 40 km or more, it is likely that lightning at Venus, if it exists, is primarily cloud-to-cloud. Terrestrial cloud-to-ground lightning is generally more intense (by about 10 to 20 dB) than cloud-to-cloud lightning, so it is possible that the absence of impulsive high-frequency radio signals during the Venus fly-bys could be owing to the dominance of very weak cloud-to-cloud lightning at Venus. The detection threshold imposed by the RPWS measurements, however, (see Fig. 4) imposes severe constraints on the intensity of any cloud-to-cloud lightning that might exist on Venus; such lightning would thus have to be much weaker than on Earth. As the previous reports of impulsive low-frequency (100 Hz to 80 kHz) radio signals by the Venera and Pioneer-Venus spacecraft indicate that some type of electrical activity is taking place in the atmosphere of Venus^{1–5}, we wonder what type of electrical activity might be involved. Because the ionospheric plasma effectively blocks these low frequency signals from reaching Cassini, we can draw no definitive conclusion about the presence or absence of such signals. However, rough comparisons of the reported Venera intensities² with the intensity limits imposed by the Cassini RPWS indicate that the spectrum must decline much

Because of the logarithmic response of the receiver, the detection threshold rises steeply for pulse durations shorter than the integration time. The detection threshold for the millisecond mode depends somewhat on the pulse shape and is difficult to quantify for very short pulses (less than 1 millisecond). This uncertainty is indicated by the shaded region. Also shown are the energy spectral densities, in units of J Hz^{-1} , of the terrestrial lightning detected during the Earth fly-by. These energy spectral densities are plotted as a function of the observed pulse duration, which are at multiples of the 21 ms integration/sampling time used during the Earth fly-by. To indicate the amplitude distribution the points have been stacked to the right whenever multiple points occur at a given energy spectral density.

more steeply than the typical $1/f^2$ spectrum of terrestrial lightning. Such a steep spectrum would indicate a relatively slow rise time for the discharge, possibly similar to the cloud-to-ionosphere discharges (sprites) that were recently discovered on Earth^{19,20}. □

Received 10 July; accepted 16 November 2000.

1. Ksanfomality, L. V. Lightning in Venus' cloud layer. *Cosmic Res.* **17**, 747–762 (1979).
2. Ksanfomality, L. V., Scarf, F. L. & Taylor, W. W. L. in *Venus* (eds Hunten, D. M., Colin, L., Donahue, T. M. & Moroz, V. I.) 565–603 (Arizona, Tucson, 1983).
3. Taylor, W. W. L., Scarf, F. L., Russell, C. T. & Brace, L. H. Evidence for lightning on Venus. *Nature* **282**, 614–616 (1979).
4. Scarf, F. L., Taylor, W. W. L., Russell, C. T. & Brace, L. H. Lightning on Venus: Orbiter detection of whistler signals. *J. Geophys. Res.* **85**, 8158–8166 (1980).
5. Russell, C. T. Venus lightning. *Space Sci. Rev.* **55**, 317–356 (1991).
6. Gurnett, D. A. *et al.* Lightning and plasma wave observations from the Galileo Flyby of Venus. *Science* **253**, 1522–1525 (1991).
7. Hunten, D. M. Venus lightning: Pros and cons. *Adv. Space Res.* **15**, 109–112 (1995).
8. Grebowsky, J. M. in *Venus II* (eds Bougher, S. W., Hunten, D. M. & Phillips, R. J.) 125–157 (Arizona, Tucson, 1997).
9. Krasnopolsky, V. A. Lightnings and nitric oxide on Venus. *Planet. Space Sci.* **31**, 1363–1369 (1983).
10. Hansel, S., Wells, S. A. & Hunten, D. M. Optical detection of lightning on Venus. *Icarus* **117**, 345–351 (1995).
11. Borucki, W. J., Dyer, J. W., Thomas, G. Z., Jordon, J. C. & Comstock, D. A. Optical search for lightning on Venus. *Geophys. Res. Lett.* **8**, 233–236 (1981).
12. Sagdeev, R. Z. *et al.* Overview of VEGA Venus balloon in situ meteorological measurements. *Science* **231**, 1411–1414 (1986).
13. Taylor, H. A. Jr, Cloutier, P. A. & Zheng, Z. Venus "lightning" signals reinterpreted as in situ plasma noise. *J. Geophys. Res.* **92**, 9907–9919 (1987).
14. Gurnett, D. A. *et al.* The Cassini radio and plasma wave investigation. *Space Sci. Rev.* (submitted).
15. Uman, M. A. *Lightning* 272 (Dover, New York, 1969).
16. Pierce, E. T. in *Lightning Vol. 1, Physics of Lightning* (ed. Golde, R. H.) 351–384 (Academic, London, 1977).
17. Brace, L. H. *et al.* in *Venus* (eds Hunten, D. M., Colin, L., Donahue, T. M. & Moroz, V. I.) 779–872 (Univ. Arizona Press, Tucson, 1983).
18. Brown, L. W. The galactic radio spectrum between 130 and 2600 kHz. *Astrophys. J.* **180**, 359–370 (1973).
19. Sentman, D. D., Wescott, E. M., Osborne, D. L., Hampton, D. L. & Heavner, M. J. Preliminary results from the Sprites 94 aircraft campaign, 1. Red sprites. *Geophys. Res. Lett.* **22**, 1205–1208 (1995).
20. Wescott, E. M., Sentman, D. D., Osborne, D. L., Hampton, D. L. & Heavner, M. J. Preliminary results from the Sprites 94 aircraft campaign, 2. Blue Jets. *Geophys. Res. Lett.* **22**, 1209–1212 (1995).

Acknowledgements

The research at the University of Iowa was supported by NASA through the Jet Propulsion Laboratory.

Correspondence and requests for materials should be addressed to D.A. (e-mail: donald-gurnett@uiowa.edu).

Tripling the capacity of wireless communications using electromagnetic polarization

Michael R. Andrews*, Partha P. Mitra* & Robert deCarvalho*†

* Bell Labs, Lucent Technologies, Murray Hill, New Jersey 07974, USA

† Harvard University, Cambridge, Massachusetts 02138, USA

Wireless communications are a fundamental part of modern information infrastructure. But wireless bandwidth is costly¹, prompting a close examination of the data channels available using electromagnetic waves. Classically, radio communications have relied on one channel per frequency, although it is well understood that the two polarization states of planar waves² allow two distinct information channels; techniques such as ‘polarization diversity’ already take advantage of this³. Recent work^{4–7} has shown that environments with scattering, such as urban areas or indoors, also possess independent spatial channels that can be used to enhance capacity greatly. In either case, the relevant signal processing techniques come under the heading of ‘multiple-input/multiple-output’ communications, because multiple antennae are required to access the polarization or spatial channels. Here we show that, in a scattering environment, an extra factor of three in channel capacity can be obtained, relative to the conventional limit using dual-polarized radio signals. The extra capacity arises because there are six distinguishable electric and magnetic states of polarization at a given point, rather than two as is usually assumed.

In Fig. 1, we show why the presence of a scattering environment violates our intuitive notion of there being only two polarization degrees of freedom for electromagnetic radiation. This situation arises because in free space, radiated electric and magnetic fields are constrained to be perpendicular to one another and to the direction of propagation². Thus, once the direction of propagation is fixed, only two degrees of freedom remain which are typically referred to as either horizontal or vertical (linear) polarizations. In the presence of a reflecting surface, however, multiple paths are possible between the two points. Although the wave propagating along the direct path

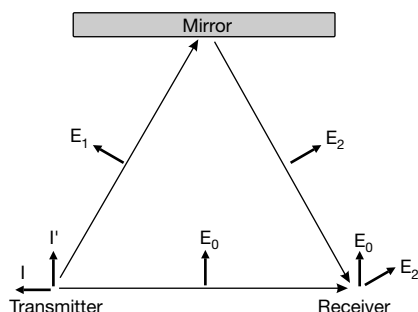


Figure 1 Communicating with electric fields in the presence of a mirror. In the two-dimensional plane of the figure, orthogonal current dipoles at the transmitter, I and I' (left), control two degrees of electric field freedom at the receiver, E_0 and E_2 (right). The direction perpendicular to the plane contributes a third degree of freedom (E_0 , E_1 , and E_2 are transverse fields on the three paths shown). Normally, due to transverse propagation of electromagnetic waves, there would be no longitudinal electric field component at the receiver. Thus, as is drawn, without the mirror the current dipole labelled I would not produce electric fields at the receiver. However, the alternative propagation path shown causes the component E_2 to appear at the receiver, which has non-zero longitudinal projection when referred to the non-reflected path.

cannot have an electric field component parallel to that path, the wave propagating along the reflected path can contribute such a component to the field at the receiver (it is effectively a longitudinal component when referred to the line-of-sight). Thus, by using appropriate transmit antennae we can propagate waves which cause independent fluctuations in all three electric field components. The electromagnetic polarization is no longer constrained to be perpendicular to the line-of-sight⁸. The presence of a single reflecting surface allows the use of three channels of electric-field polarization for wireless communication.

In order to make our statements more precise, we introduce H , a 6×6 matrix relating the electric ($\mathbf{E}$) and magnetic ($\mathbf{B}$) fields measured at a point $\mathbf{r}$, owing to idealized oscillating electric ($\mathbf{p}$) and magnetic ($\mathbf{m}$) dipole moments, produced by transmitting antennae at point $\mathbf{r}'$

$$\begin{bmatrix} \mathbf{E}(\mathbf{r}) \\ c\mathbf{B}(\mathbf{r}) \end{bmatrix} = -H(\mathbf{k}, \mathbf{r} - \mathbf{r}') \begin{bmatrix} c\mathbf{p} \\ \mathbf{m} \end{bmatrix} \quad (1)$$

where in free space and in the far-field, H is given by a compact version of the standard formulas that describe the fields radiated by oscillating electric and magnetic dipoles² (in SI units):

$$H_0(\mathbf{k}, \mathbf{r}) = \frac{|\mathbf{k}|^3}{\epsilon_0 c} \frac{e^{i\mathbf{k} \cdot \mathbf{r}}}{4\pi\mathbf{k} \cdot \mathbf{r}} \begin{bmatrix} J^2(\hat{\mathbf{r}}) & J(\hat{\mathbf{r}}) \\ -J(\hat{\mathbf{r}}) & J^2(\hat{\mathbf{r}}) \end{bmatrix} \quad (2)$$

Here $\mathbf{k}$ is the wave vector ($|\mathbf{k}| = \omega/c = 2\pi/\lambda$) and $1/\epsilon_0 c \approx 377 \Omega$ is the impedance of free space (c is the speed of light). $J(\hat{\mathbf{r}})$ is a 3×3 matrix given by $J_{ij}(\hat{\mathbf{r}}) = \sum_k \epsilon_{ikj} \hat{\mathbf{r}}_k$, defined so that $J(\hat{\mathbf{r}})\mathbf{p} = \hat{\mathbf{r}} \times \mathbf{p}$ (with $\hat{\mathbf{r}} = \mathbf{r}/|\mathbf{r}|$).

The communication channel associated with equation (1) includes additive noise measured by the receiver. For simplicity, we make the canonical assumption that its components are uncorrelated gaussian white noise with equal variance, and that communication takes place over a sufficiently narrow bandwidth that H has negligible frequency dependence (‘flat fading’). For a time-independent H , the rate at which information can be transferred

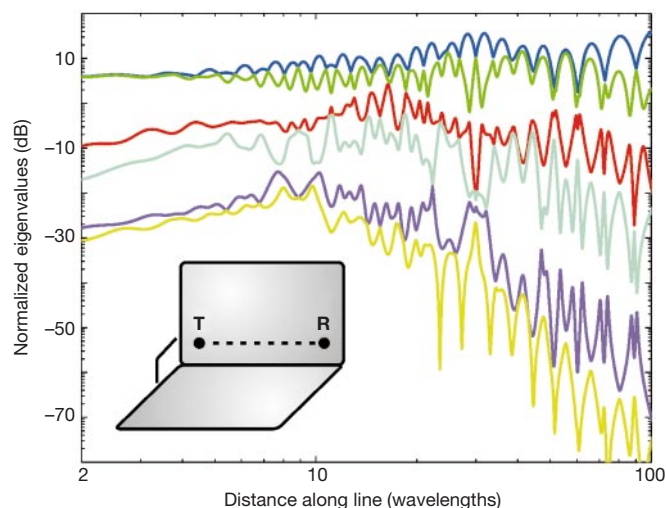


Figure 2 Eigenvalues of $HH^\dagger$ in a two-mirror idealized scattering environment. Data were obtained by simulating the receiver (R) at a variable distance from the transmitter (T) along the line indicated by the inset. The line joins the points $(x, y, z) \approx (9.9, 7.7, 10.5)\lambda$ and $(15.1, 109.8, 8.1)\lambda$, while the mirrors are in the planes $x = 0$ and $z = 0$. Note that $\text{rank}(H) = 6$ (that is, there are no eigenvalues equal to zero), indicating a threefold capacity increase over what is possible in free space with no scattering. Eigenvalues are normalized to those that would be obtained in free space: $(|\mathbf{k}|^2/4\pi\epsilon_0|\mathbf{r}|c)^2$. (At large distances the reflections are glancing, which leads to widely spaced eigenvalues). See equations (1) and (2).

between an n -antennae transmitter/receiver pair ($n = 6$ in equations (1) and (2) above) is characterized by the quantity

$$M(H) = \log_2 \det[I + (\rho/n)HH^\dagger] \quad (3)$$

where $M(H)$, in bits per second per hertz, is the mutual information between the transmitter and receiver when the transmitted signals are uncorrelated white gaussian stochastic processes with equal power. We use units such that the noise components have unit variance, allowing us to denote both total power and signal-to-noise ratio by the same symbol ρ . The significance of $M(H)$ is that for a fluctuating H , under appropriate conditions, the capacity C of the channel is given by the expectation of $M(H)$ taken over the probability distribution of H . The conditions are that H is known by the receiver but not by the transmitter (as would be the case if transmitted pilot signals help the receiver calculate H). We note that if the channel is known to the transmitter, the capacity will be higher and is given by the “water filling” solution where power is unevenly divided among transmitting antennae⁶. Fluctuations in H may be caused by a fluctuating environment or by movement of the antennae.

Clearly, the capacity depends on the rank of H . It follows by inspection that at large signal-to-noise ratio ρ , $M(H)$ (and therefore C) tends to the value $m \log_2 \rho$, where $m = \text{rank}(H)$. We note that the large ρ limit is taken for fixed H . It is easy to verify that the rank of H_0 (see equation (2)) is two, which corresponds to our intuitive notion of there being only two polarization degrees of freedom in free space. We now formally define the number of polarization channels as $\text{rank}(H)$. Our basic observation here may be summarized by the statement that it is possible to have $\text{rank}(H) = 6$ in an environment with scattering, with a concomitant increase in the capacity of the wireless communication channel. We note that six independent signals have to be transmitted in order to take advantage of this increased capacity.

A simple geometry which shows that we would expect $\text{rank}(H) = 6$ in scattering environments consists of two perfectly conducting planes (mirrors) at right angles to each other (inset of Fig. 2). The matrix H (equation (1)) in this environment may be computed by summing over free-space contributions like H_0 (equation (2)) corresponding to the actual transmitter and its images in the two mirrors (two single and one double reflection). We plot the eigenvalues of $HH^\dagger$ for the transmitter and receiver placed along a given line in the simulated environment (Fig. 2); we note that all six eigenvalues are non-zero and hence $\text{rank}(H) = 6$, signifying a threefold increase in capacity over what would be possible in free space. An alternative way of seeing that all six components of the electromagnetic field can fluctuate independently is to consider their covariance at a point due to an isotropic distribution of plane waves. It is easy to verify in this case that all



Figure 3 Tri-polarized antenna (tripole). The composite structure is composed of three orthogonal sleeves¹² with their feed points collocated at the centre of the image. Each has two variable length segments which are adjusted to efficiently couple radiation to its transmission line. The three antennae together can create an arbitrary vector electric dipole moment at the transmitter, as well as function as detectors of the vector electric field at the receiver.

off-diagonal covariances vanish, so that no vector component can be predicted from any other.

Apart from increasing the number of degrees of freedom in wireless electromagnetic communications, the existence of six independent channels as indicated by the full rank of H also implies improved fading performance of a receiver sensitive to polarization degrees of freedom. As the environment fluctuates, or as the receiver moves through space, the measured electric and magnetic fields also fluctuate. For an antenna sensitive to just a single field component the amplitude can occasionally come close to zero (fading); however, it is quite unlikely that all six vector components will vanish simultaneously. Thus, by taking full advantage of the receive diversity that is offered by polarization (which we noted is already done for two polarization states³) the fading can be greatly ameliorated. In contrast with earlier proposals^{4–7}, where scattering leads to greater spatial diversity, the subject of this Letter involves measurements theoretically localizable at only one spatial point, thus leading to more compact antenna designs for similar capacity gains. At one spatial point it is possible to measure two three-dimensional vector fields (electric and magnetic); we do not consider, however, measurements of spatial field derivatives, which are independent degrees of freedom and in principle could lead to further increases in capacity beyond the factor of three considered here. Completely variable vector electric and magnetic fields seem only to have been discussed previously in direction-finding applications^{9–11}.

Our experimental work demonstrates the increase in capacity originating from the three electric polarization states in a scattering environment. Dubbed a “tripole”, the antenna system consists of three standard sleeves¹² arranged orthogonally as shown in Fig. 3. Each antenna element is roughly a half wavelength long (~ 17 cm), and was trimmed for efficient resonant operation at 880 MHz by adjusting variable length sections. Each has good coupling of power to radiated modes with less than 5% power reflected back to the transmitting source. As the three elements are orthogonal and cross close to their feed points, coupling among them is minimized (less than -25 dB of power) and we find the radiation pattern of each element in the tripole to be similar to that of an isolated sleeve including deep nodes (-25 dB of power) along the axes. The three antennae together can create an arbitrary vector electric dipole

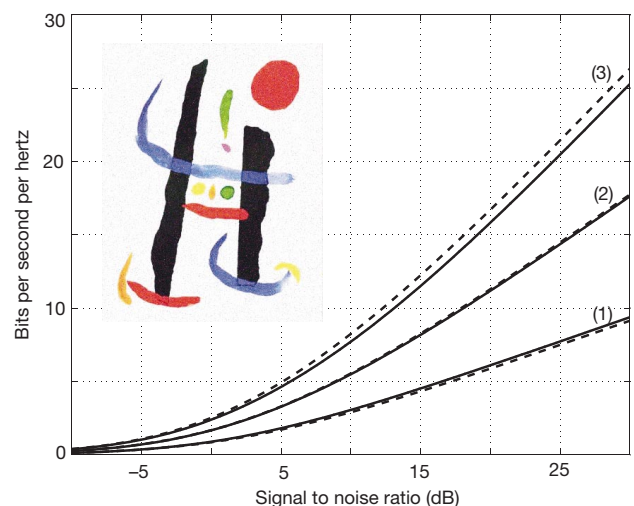


Figure 4 Capacity of the tripole antenna. Shown are experimentally obtained capacity estimates for one (lower solid line), two and three (upper solid lines) orthogonal sleeve antennae (Fig. 3), calculated from measured transfer matrices H . Dotted lines are from the random matrix-based theory⁵. Inset, here we show a capacity increase through the transfer of actual data. The reconstructed colour image above (Joan Miro, *A toute épreuve*) was transmitted as three distinct red, green, and blue monochrome subfields on each of the three orthogonally oriented antennae of Fig. 3 (that is, red was transmitted on $\hat{x}$ polarization, green on $\hat{y}$, and blue on $\hat{z}$).

moment at the transmitter, as well as function as detectors of the vector electric field at the receiver. As for the magnetic polarization components, we assume that three more resonant antenna structures, perhaps mutually orthogonal magnetic-sensing 'loops', could be designed to be collocated with the three orthogonal (electric-sensing) sleeves. Although the engineering details associated with magnetic-sensing antennae differ from those of electric-sensing ones, these differences have been successfully dealt with before^{12,13}.

Having constructed two such tri-polarized antennae, demonstrating their potential for increased capacity in a particular environment involves measuring the rank of the 3×3 electric transfer matrix H (that is, equation (1)) where contributions from $\mathbf{B}$ and $\mathbf{m}$ are ignored). A variety of pilot signals may be used to measure the transfer matrix H , but we chose simple sinusoids (tones) for purposes of characterization. Let α and β each denote one of the three indices of the transmit and receive antennae polarizations. The experimental procedure was to choose three distinct tones ω_β within a narrow frequency band (experimentally, it is found that H (equation (1)) varies negligibly over bands of 20 kHz) and transmit each on its own antenna element. All three frequencies were then observed on each of the three antenna elements at the receiver. The transfer matrix element $H_{\alpha\beta}$ could then be read off as the complex amplitude at frequency ω_β (transmitted on antenna β) received on antenna α .

The results are illustrated in the graph of Fig. 4, and show capacities inferred from the measured matrices. The upper curve is the "tripole" capacity (3×3 matrix), whereas the lower and middle curves are for 1×1 (scalar) and 2×2 (dual-polarized) submatrices, respectively. In the scalar and dual-polarized cases we present capacities averaged over the nine possible choices for antennae at the transmitter and receiver. (There are three possible ways of choosing either one or two antennae from three, and combinations of transmit and receive elements yields 3×3 or nine possibilities.) It is clear from the figure that use of polarization adds to the capacity, and does so in agreement with theory⁶. These data were taken in a large cafeteria (about 500 m², with 10-m high ceilings) with the transmitter and receiver around a corner from one another; each matrix was measured with the receiver at a different position but always about 25 m from the transmitter.

In order to provide an illustrative visual example of the expansion of channel capacity, we wirelessly communicated a colour image with colour represented as usual by a three-dimensional vector of red, green and blue components. These three monochrome image components of the original full-colour image were transmitted separately by each of the three transmit antennae, along with pilot signals allowing for measurement of the 3×3 transfer matrix H . The inset of Fig. 4 shows the recovered image (Joan Miro, *A toute épreuve*).

From these studies it is clear that electromagnetic polarization can play an important role in wireless communication. We have demonstrated experimentally the capacity increase that comes from exploiting the three-dimensional electric field vector (Fig. 4), and have theoretically shown the benefit arising from the use of magnetic degrees of freedom as well (Fig. 2). There will undoubtedly be practical issues to be considered in using a sixfold polarized antennae system, including the efficiency and cross-coupling of antennae extracting all the electromagnetic degrees of freedom near a point, as well as issues such as the statistics and environmental fluctuations that affect H . Nevertheless, the capacity increase is real and necessitates a re-evaluation of the use of polarization in wireless communication using electromagnetic waves. □

Received 4 July; accepted 26 October 2000.

- Andrews, E. L. \$50 billion for German wireless licenses. *The New York Times* 18 August (2000).
- Jackson, J. D. *Classical Electrodynamics* (John Wiley, New York, 1975).
- Singer, A. Space vs. polarization diversity. *Wireless Rev.* **15**, 164–166 (1998).
- Winters, J. H. On the capacity of radio communication systems with diversity in a rayleigh fading environment. *IEEE J. Selected Areas Commun.* Vol. SAC-5, 871–878 (1987).

- Foschini, G. J. & Gans, M. J. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Pers. Commun.* **6**, 311–335 (1998).
- Telatar, I. E. Capacity of multi-antenna gaussian channels. *Eur. Trans. Telecommun.* **10**, 585–595 (1999).
- Moustakas, A. L., Baranger, H. U., Balents, L., Sengupta, A. M. & Simon, S. Communication through a diffusive medium: Coherence and capacity. *Science* **287**, 287–290 (2000).
- deCarvalho, R., Mitra, P. P. & Andrews, M. R. in *National Radio Science Meeting 84* (International Union of Radio Science, National Academy of Sciences, 2000).
- Morgan, M. & Evans, W. Synthesis and analysis of elliptic polarization loci in terms of space-quadrature sinusoidal components. *Proc. IRE* **39**, 552–556 (1951).
- Hatke, G. F. in *Twenty-seventh Asilomar Conference on Signals, Systems & Computers* 1365–1369 (IEEE Computer Society Press, Los Alamitos, California, 1993).
- Afraimovich, E. L., Chernukhov, V. V., Kobzar, V. A. & Palamarchouk, K. S. Determining polarization parameters and angles of arrival of hf radio signals using three mutually orthogonal antennas. *Radio Sci.* **34**, 1217–1225 (1999).
- Kraus, J. D. *Antennas* 2nd edn 725–726 (McGraw Hill, Boston, 1988).
- Lee, W. C. Y. Theoretical and experimental study of the properties of the signal from an energy-density mobile-radio antenna. *IEEE Trans.* **1**, 25–32 (1967).

Acknowledgements

This work is supported by internal funding at Bell Labs, Lucent Technologies. We are grateful to W. M. MacDonald for his assistance in characterizing our antenna's radiation patterns, and to M. J. Gans for his insights and discussions concerning magnetic polarization.

Correspondence and requests for materials should be addressed to M.R.A. (e-mail: mikea@bell-labs.com).

Relationship between structural order and the anomalies of liquid water

Jeffrey R. Errington & Pablo G. Debenedetti

Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544-5263, USA

In contrast to crystalline solids—for which a precise framework exists for describing structure¹—quantifying structural order in liquids and glasses has proved more difficult because even though such systems possess short-range order, they lack long-range crystalline order. Some progress has been made using model systems of hard spheres^{2,3}, but it remains difficult to describe accurately liquids such as water, where directional attractions (hydrogen bonds) combine with short-range repulsions to

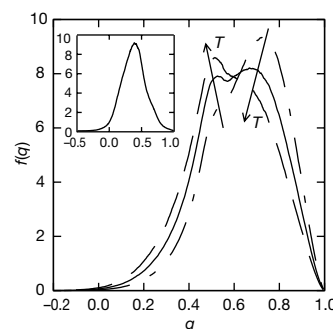


Figure 1 The effect of temperature on the distribution of the orientational order, q , at a density $\rho = 1.2 \text{ g cm}^{-3}$. Three temperatures T are considered; 320 K (dashed line), 280 K (solid line) and 240 K (dot-dashed line). The fraction of molecules with q -values between $q + dq/2$ and $q - dq/2$ is $f(q) dq$. The area under each curve is normalized to unity. Arrows indicate the effect of increasing temperature. The possible range of q for a molecule is $-3 \leq q \leq 1$; the average value for a collection of molecules spans the range $0 \leq \langle q \rangle \leq 1$. Inset, the q -distribution for a Lennard-Jones system at a density $\rho^* = N\sigma^3/V = 1.0$ and temperature $T^* = k_B T/\epsilon = 2.0$, where ϵ and σ are the energy and size parameters, and k_B is Boltzmann's constant. For the Lennard-Jones fluid, $f(q)$ is virtually unchanged throughout the liquid region of the phase diagram.

determine the relative orientation of neighbouring molecules as well as their instantaneous separation. This difficulty is particularly relevant when discussing the anomalous kinetic and thermodynamic properties of water, which have long been interpreted qualitatively in terms of underlying structural causes. Here we attempt to gain a quantitative understanding of these structure–property relationships through the study of translational^{2,3} and orientational⁴ order in a model⁵ of water. Using molecular dynamics simulations, we identify a structurally anomalous region—bounded by loci of maximum orientational order (at low densities) and minimum translational order (at high densities)—in which order decreases on compression, and where orientational and translational order are strongly coupled. This region encloses the entire range of temperatures and densities for which the anomalous diffusivity^{6–9} and thermal expansion coefficient¹⁰ of water are observed, and enables us to quantify the degree of structural order needed for these anomalies to occur. We also find that these structural, kinetic and thermodynamic anomalies constitute a cascade: they occur consecutively as the degree of order is increased.

We introduce two measures of order in water^{11–14}. The translational order parameter³ t measures the tendency of pairs of molecules to adopt preferential separations (see Methods). It vanishes for an ideal gas, and is large for a crystal. The orientational order parameter⁴ q measures the extent to which a molecule and its four nearest neighbours adopt a tetrahedral arrangement¹⁵, such as exists in hexagonal ice (I_h) (see Methods). It vanishes for an ideal gas, and equals 1 in a perfectly tetrahedral arrangement. Figure 1 shows the calculated distribution of q -values at different temperatures and fixed density (see Methods for details of the molecular dynamics simulations). The bimodal distribution suggests that the transient arrangement^{14,16,17} adopted by a molecule and its four nearest neighbours can be described as being predominantly structured and ‘ice’-like (high- q peak) or unstructured (low- q peak)^{18,19}. This should not be interpreted as implying the existence of two fixed types of energetically favoured arrangements.

Figure 2 shows the calculated evolution of t and q on compression. At 260 K, starting from $\rho = 0.85 \text{ g cm}^{-3}$, both t and q increase with density. At point A, water attains the maximum orientational order, and the hydrogen-bond network is at its most structured. Further compression leads to a decrease in q . Translational order is strongly coupled to orientational order under these conditions, causing t also to decrease upon compression. This means that hydrogen-bonding determines both the mutual orientation and the separation between molecules. Increasing the temperature weakens the coupling between translational and orientational

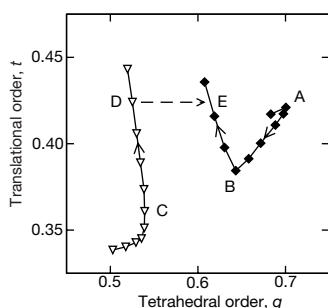


Figure 2 The path traversed in order-parameter space as liquid water is compressed isothermally at two different temperatures. Filled diamonds, $T = 260 \text{ K}$; open triangles, $T = 400 \text{ K}$. The arrows indicate the direction of increasing density, which varies from 0.85 to 1.3 g cm^{-3} ($T = 260 \text{ K}$) and from 0.8 to 1.3 g cm^{-3} ($T = 400 \text{ K}$), each symbol representing an increment of 0.05 g cm^{-3} . A and C are states of maximum orientational order at the respective temperatures. B is a state of minimum translational order. Along path D–E, t and q are independently variable, as one of them varies while the other is kept fixed.

order, and the decrease in q upon compression does not cause a corresponding decrease in t (400 K isotherm). A maximum in orientational order is attained at point C in Fig. 2. At this temperature, however, compression always causes translational order to increase, as in the hard-sphere liquid^{2,3}.

In order to investigate the relationship between molecular order and dynamics, we show in Fig. 3 the calculated dependence of the diffusion coefficient D upon density and temperature. Diffusivity maxima in water are well-documented experimentally^{7,8} and by computer simulation⁹. Diffusivity minima have been reported only in simulations⁶. The anomalous region where $dD/d\rho$ is greater than zero disappears above 300 K . Figure 4 shows the calculated loci of diffusivity minima, diffusivity maxima, q maxima (for example, point A in Fig. 2), t minima (for example, point B in Fig. 2), and density maxima²⁰ (also known as temperature of maximum density or TMD²¹). Along the TMD locus, water’s thermal expansion coefficient vanishes, and inside the region it encloses, the thermal expansion coefficient is negative. At low temperatures, the locus of maximum orientational order is very close to the locus of minimum diffusivity, with the ‘closeness’ becoming more pronounced the lower the temperature; at the lowest temperature studied, these loci

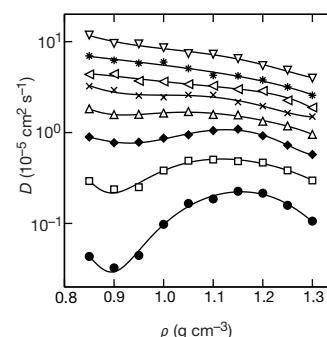


Figure 3 Density dependence of the diffusion coefficient, D , shown for eight isotherms. The curves from top to bottom correspond to temperatures T (in K) of: 400, 350, 320, 300, 280, 260, 240 and 220. The solid lines are fifth-order polynomial fits to the data, and are simply guides to the eye. The diffusion coefficient was calculated from the long-time behaviour of the mean squared displacement of the water molecules, $\langle r^2(t) \rangle$, using the Einstein relationship $6D = d\langle r^2(t) \rangle/dt$.

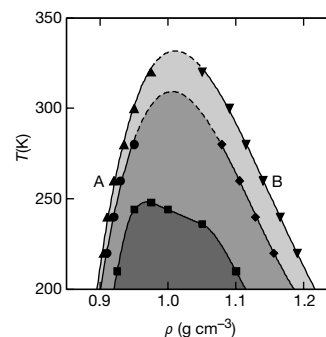


Figure 4 Relationship between loci of structural, dynamic and thermodynamic anomalies. The structurally anomalous region is bounded by the loci of q maxima (upward-pointing triangles) and t minima (downward-pointing triangles). Inside this region, water becomes more disordered when compressed, as t and q decrease with increasing density. The loci of diffusivity minima (circles) and maxima (diamonds) define the region of diffusive anomalies, where D increases with density. The thermodynamically anomalous region is defined by the TMD (squares), inside which the density increases when water is heated at constant pressure. We refer to the low-density boundary of the structurally anomalous region as a q -maximum locus, with the understanding that this applies to points such as A, which are maxima in q at which t decreases with increasing density, but not to C (Fig. 2), which lies outside the structurally anomalous region and for which the q maximum is not accompanied by a decrease in t .

are virtually within numerical error of each other. The locus of minimum translational order parallels the locus of diffusivity maxima. The TMD, and hence all states where the thermal expansion coefficient is negative, lies entirely inside the region defined by the loci of maximum q and minimum t . We refer to the region bounded by t and q extrema as structurally anomalous, because inside it orientational and translational order decrease by compression.

If the order parameters of states inside the structurally anomalous region are plotted in the (t, q) plane, we obtain, within the numerical accuracy of the simulations, a single line (Fig. 5). States outside the structurally anomalous region map onto the region of the (t, q) plane lying above this line. Therefore, all structurally anomalous states have the important property that their orientational and translational order are not independent, there being no thermodynamic path along which one can be varied while keeping the other fixed. Alternative scenarios, such as each isotherm being tangent at one point to—or coinciding over a limited range with—a common (t, q) envelope, cannot be ruled out within the numerical

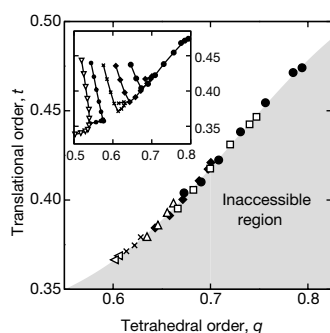


Figure 5 The line of structural anomalies in order-parameter space. The data points in the main panel correspond to state points that lie between the maximum in orientational order and the minimum in translational order along an isotherm, for example state points between A and B in Fig. 2. All state points within the structurally anomalous region bounded by triangles in Fig. 4 map onto the line shown in the main panel. Symbols for different temperatures are as in Fig. 3. Inset, five isotherms in order-parameter space that span the density range $0.85 \leq \rho \leq 1.3 \text{ g cm}^{-3}$ (0.8 to 1.3 g cm^{-3} at 400 K). The data points in the main panel represent a subset of the points displayed in the inset. This subset corresponds to thermodynamic states inside the structurally anomalous region.

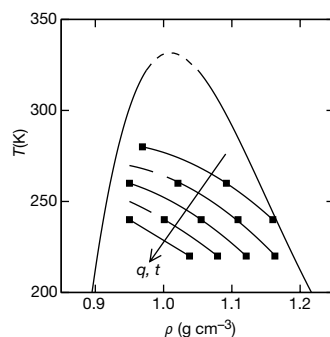


Figure 6 Isotaxis lines in the (T, ρ) plane. The q -difference between lines is constant ($\Delta q = 0.02$). The arrow indicates the direction of increasing order. The (q, t) values along the isotaxes (top to bottom) are $(0.66, 0.394)$, $(0.68, 0.405)$, $(0.70, 0.417)$, $(0.72, 0.429)$ and $(0.74, 0.442)$. The outer solid and dashed lines define the structurally anomalous region. Lines through squares are guides to the eye. The slope of isentropic lines (not shown) has the same sign as the thermal expansion coefficient. Inside the TMD, therefore, isentropes and isotaxes have negative slope. Inside the structurally anomalous region but outside the TMD, isentropes have positive slope while isotaxes have negative slope: t and q do not possess the same information content as the entropy. It is not clear how much of the missing information is related to vibrational degrees of freedom or to higher-order correlations not included in t and q .

precision of the data. Outside the structurally anomalous region orientational and translational order are independently variable, as there exist paths along which one can be varied while the other is kept fixed, such as D–E (Fig. 2.) States inside the structurally anomalous region possess the maximum orientational order consistent with their translational order. Over the range of temperatures and densities investigated, we found no states lying below the line shown in Fig. 5.

Because the line on the (t, q) plane onto which all structurally anomalous states are mapped has positive slope, it is possible to rank all such equilibrium states according to their relative order. This is done in Fig. 6, where several lines of constant t and q are shown. We call these isotaxis lines (from the Greek word *taxis*, meaning order, or arrangement). The isotaxis lines have negative slope in the (T, ρ) plane, with t and q jointly increasing in the direction of decreasing T and ρ . This reflects the fact that water becomes more structured upon cooling and expansion.

The physical picture that emerges from this work is the following. In liquid water there occurs a cascade of anomalies. Structural anomalies, whereby order decreases upon compression, occur over the broadest range of densities and temperatures. Diffusive anomalies, whereby D increases by compression, occur entirely inside the region of structural anomalies. Thermodynamic anomalies, whereby the density decreases upon cooling at constant pressure, occur entirely inside the region of diffusive anomalies. All anomalous states share the topological property that orientational and translational order are strongly coupled. Isotaxis lines rank thermodynamic states according to order. To each boundary of anomalies there corresponds a limiting tangent isotax that defines the minimum order needed for that anomaly to occur. For SPC/E water we find $q > 0.615$ for diffusive anomalies, and $q > 0.705$ for density anomalies.

As water's loci of diffusivity maxima and configurational entropy maxima coincide⁹, future investigations could usefully determine whether the loci of diffusivity minima and configurational entropy minima also correspond. This would clarify the relationship between configurational entropy^{22–24} and the structural indicators of anomalies introduced here. Liquid water possesses additional anomalies^{10,25} not discussed here—for example, increase in the isothermal compressibility²⁶ and isobaric heat capacity^{27,28} upon cooling, decrease in viscosity upon compression²⁹, a polymorphic transition^{10,19,21,25} and the possible existence of a second critical point^{10,21,25,30}. Another area for future work would be the determination of the structural limits where each component in the cascade of anomalies occurs. Our results suggest that the quantification of order may offer new insights into the properties of liquids, glasses and other condensed-phase systems. □

Methods

Order parameters

The translational order parameter t is given by³

$$t = \frac{\int_0^{\xi_c} |g(\xi) - 1| d\xi}{\xi_c}$$

where $\xi (= r\rho^{1/3})$ is the distance between the oxygen atoms of a pair of molecules, r , divided by the mean separation between a pair of molecules at the given density, $\rho^{-1/3}$; ρ denotes the density, or number of molecules per unit volume (N/V); g is the oxygen–oxygen radial distribution function; and ξ_c is a cut-off distance. We use $\xi_c = 2.843$ in this work. In an ideal gas, $g = 1$ and t vanishes. In a crystal, on the other hand, there is long-range translational order, and $g \neq 1$ over long distances. Hence t is large. Alternative definitions of translational order can be used when the underlying crystal structure is known^{2,3}. However, the stable crystal form of the SPC/E model of water is not known. The basic features that define the relationship between translational order and density, and therefore between translational and orientational order, are insensitive to the precise definition of t (refs 2, 3). The orientational order parameter is given by

$$q = 1 - \frac{3}{8} \sum_{j=1}^3 \sum_{k=j+1}^4 \left(\cos \psi_{jk} + \frac{1}{3} \right)^2$$

where ψ_{jk} is the angle formed by the lines joining the oxygen atom of a given molecule and those of its nearest neighbours j and k (≤ 4). This definition is a rescaled version of the parameter introduced in ref. 4, in such a way that the average value of q varies between 0 (in an ideal gas) and 1 (in a perfect tetrahedral network). If a molecule is located at the centre of a regular tetrahedron whose vertices are occupied by its four nearest neighbours, $\cos\psi_{jk} = -1/3$. Thus, in a perfect tetrahedral network, $q = 1$. If, on the other hand, the mutual arrangement of molecules is random, as in an ideal gas, the six angles associated with the central molecule are independent, and the mean value of q vanishes:

$$\langle q \rangle = 1 - \frac{9}{8} \int_0^\pi \left(\cos\psi + \frac{1}{3} \right)^2 \sin\psi d\psi = 0$$

Alternative measures of order can be used. They do not change the general picture shown in Fig. 4. We have tested the 'tetrahedrality parameter' of ref. 31 as an alternative measure of orientational order, and the two-body excess entropy of ref. 32 as an alternative measure of translational order². Isotherms of these order parameters exhibit extrema with respect to density that are within 1.7% of those of q and t . The extrema loci define a structurally anomalous region almost identical to the one shown in Fig. 4. In particular, it lies entirely outside the region of diffusive anomalies, confirming that structural anomalies occur over the broadest range of densities and temperatures and anticipate the occurrence of transport and thermodynamic anomalies.

Molecular dynamics simulations

Our results are based on constant temperature and density molecular dynamics simulations³³ of 256 particles interacting via the shifted-force SPC/E potential⁵ in a cubic box with periodic boundary conditions. The Ewald summation³³ was used to account for electrostatic interactions. The Lennard-Jones potential and real-space part of the Ewald summation were cut and linearly shifted to bring the energy and force to zero at a distance of 7.9 Å. The temperature was maintained using a Berendsen rescaling of the velocities with a relaxation time of 0.5 ps, and the equations of motion were solved using the velocity-Verlet and the RATTLE algorithm with a time step of 2 fs (ref. 33). For each of the densities examined, initial configurations were generated randomly and equilibrated using Monte Carlo techniques at the highest temperature considered in this study, 400 K. Molecular dynamics simulations were started using the final particle positions of a previous simulation at a state point as near as possible to the new one. Each system was equilibrated for between 0.5 and 15 ns, depending on the state point, followed by a production phase (of length greater than or equal to the equilibration phase), during which the properties of interest were sampled every 40 ps.

Location of diffusivity extrema

Isochores of the $T \leq 300$ K data were fitted to the power law $D(T) = D_0(T/T_0 - 1)^\gamma$, where D_0 , T_0 and γ are density-dependent adjustable parameters. Subsequently, the diffusivity minima were located from spline fits to isotherms generated by the power-law functions, and the diffusivity maxima were obtained from fifth-order polynomial fits to the isotherms. The temperatures of maximum density (squares in Fig. 4) were obtained from fourth-order polynomial fits to isochors of pressure as a function of temperature. The positions of the extrema in orientational and translational order were extracted from spline fits to isotherms of the respective order parameter as a function of density.

Received 12 September; accepted 14 November 2000.

1. Ashcroft, N. W. & Mermin, N. D. *Solid State Physics* (Saunders College Publishing, Fort Worth, 1976).
2. Torquato, S., Truskett, T. M. & Debenedetti, P. G. Is random close packing of spheres well defined? *Phys. Rev. Lett.* **84**, 2064–2067 (2000).
3. Truskett, T. M., Torquato, S. & Debenedetti, P. G. Towards a quantification of disorder in materials. Distinguishing equilibrium and glassy sphere packings. *Phys. Rev. E* **62**, 993–1001 (2000).
4. Chau, P.-L. & Hardwick, A. J. A new order parameter for tetrahedral configurations. *Mol. Phys.* **93**, 511–518 (1998).
5. Berendsen, H. J. C., Grigera, R. J. & Stroatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
6. Ruocco, G., Sampoli, M., Torcini, A. & Vallauri, R. Molecular dynamics results for stretched water. *J. Chem. Phys.* **99**, 8095–8104 (1993).
7. Prielmeier, F. X., Lang, E. W., Speedy, R. J. & Lüdemann, H.-D. Diffusion in supercooled water to 300 MPa. *Phys. Rev. Lett.* **59**, 1128–1131 (1987).
8. Angell, C. A., Finch, E. D., Woolf, L. A. & Bach, P. Spin-echo diffusion coefficients of water to 2380 bar and -20°C . *J. Chem. Phys.* **65**, 3063–3066 (1976).
9. Scala, A., Starr, F. W., La Nave, E., Sciortino, F. & Stanley, H. E. Configurational entropy and diffusivity of supercooled water. *Nature* **406**, 166–169 (2000).
10. Debenedetti, P. G. *Metastable Liquids. Concepts and Principles* (Princeton University Press, Princeton, 1996).
11. Tanaka, H. Simple physical explanation of the unusual thermodynamic behavior of liquid water. *Phys. Rev. Lett.* **80**, 5750–5753 (1998).
12. Tanaka, H. Two-order-parameter description of liquids: critical phenomena and phase separation in supercooled liquids. *J. Phys. Condens. Matter* **11**, L159–L168 (1999).
13. Tanaka, H. Simple physical model of liquid water. *J. Chem. Phys.* **112**, 799–809 (2000).
14. Shiratani, E. & Sasaki, M. Molecular scale precursor of the liquid-liquid phase transition of water. *J. Chem. Phys.* **108**, 3264–3276 (1998).
15. Paschek, D. & Geiger, A. Simulation study on the diffusion motion in deeply supercooled water. *J. Phys. Chem. B* **103**, 4139–4146 (1999).
16. Sciortino, F., Poole, P. H., Stanley, H. E. & Havlin, S. Lifetime of the bond network and gel-like anomalies in supercooled water. *Phys. Rev. Lett.* **64**, 1686–1689 (1990).
17. Sciortino, F., Geiger, A. & Stanley, H. E. Effect of defects on molecular mobility in liquid water. *Nature* **354**, 218–221 (1991).

18. Soper, A. K. & Ricci, M. A. Structures of high-density and low-density water. *Phys. Rev. Lett.* **84**, 2881–2884 (2000).
19. Bellissent-Funel, M.-C. Is there a liquid-liquid phase transition in supercooled water? *Europhys. Lett.* **42**, 161–166 (1998).
20. Angell, C. A. & Kanno, H. Density maxima in high-pressure supercooled water and liquid silicon dioxide. *Science* **193**, 1121–1122 (1976).
21. Poole, P. H., Sciortino, F., Essman, U. & Stanley, H. E. Phase behavior of metastable water. *Nature* **360**, 324–328 (1992).
22. Sciortino, F., Kob, W. & Tartaglia, P. Inherent structure entropy of supercooled liquids. *Phys. Rev. Lett.* **83**, 3214–3217 (1999).
23. Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **267**, 1935–1939 (1995).
24. Goldstein, M. Viscous liquids and the glass transition. A potential energy barrier picture. *J. Chem. Phys.* **51**, 3728–3739 (1969).
25. Mishima, O. & Stanley, H. E. The relationship between liquid, supercooled and glassy water. *Nature* **396**, 329–335 (1998).
26. Speedy, R. J. & Angell, C. A. Isothermal compressibility of supercooled water and evidence for a thermodynamic singularity at -45°C . *J. Chem. Phys.* **65**, 851–858 (1976).
27. Angell, C. A., Oguni, M. & Sichina, W. J. Heat capacity of water at extremes of supercooling and superheating. *J. Phys. Chem.* **86**, 998–1002 (1982).
28. Tombari, E., Ferrari, C. & Salvetti, G. Heat capacity anomaly in a large sample of supercooled water. *Chem. Phys. Lett.* **300**, 749–751 (1999).
29. DeFries, T. & Jonas, J. Pressure dependence of NMR proton spin-lattice relaxation times and shear viscosity in liquid water in the temperature range -15 to 10°C . *J. Chem. Phys.* **66**, 896–901 (1977).
30. Mishima, O. & Stanley, H. E. Decompression-induced melting of ice IV and the liquid-liquid transition in water. *Nature* **392**, 164–168 (1998).
31. Naberukhin, Y. I., Voloshin, V. P. & Medvedev, N. M. Geometrical analysis of the structure of simple liquids: percolation approach. *Mol. Phys.* **73**, 917–936 (1991).
32. Nettleton, R. E. & Green, M. S. Expression in terms of molecular distribution functions for the entropy density in an infinite system. *J. Chem. Phys.* **29**, 1365–1370 (1958).
33. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1990).

Acknowledgements

We thank T.M. Truskett and S. Torquato for discussions. This work was supported by the US Department of Energy and Unilever Research.

Correspondence and requests for materials should be addressed to P.G.D. (e-mail: pdebene@princeton.edu).

Transparent nematic phase in a liquid-crystal-based microemulsion

Jun Yamamoto* & Hajime Tanaka

Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

Complex fluids^{1,2} are usually produced by mixing together several distinct components, the interactions between which can give rise to unusual optical and rheological properties of the system as a whole. For example, the properties of microemulsions (composed of water, oil and surfactants) are determined by the microscopic structural organization of the fluid that occurs owing to phase separation of the component elements. Here we investigate the effect of introducing an additional organizing factor into such a fluid system, by replacing the oil component of a conventional water-in-oil microemulsion with an intrinsically anisotropic fluid—a nematic liquid crystal. As with the conventional case, the fluid phase-separates into an emulsion of water microdroplets (stabilized by the surfactant as inverse micelles) dispersed in the 'oil' phase. But the properties are further influenced by a significant directional coupling between the liquid-crystal molecules and the surfactant tails that emerge (essentially radially) from the micelles. The result is a modified bulk-liquid crystal that is an ordered nematic at the mesoscopic level, but which does not exhibit the strong light scattering generally associated with bulk nematic order²: the bulk material here is essentially isotropic and thus transparent.

* Present address: Yokoyama Nano-structured Liquid-crystal Project, Tsukuba 300-2635, Japan.

Complex fluids^{1,2} are generally multi-component mixtures. They sometimes possess mixed physical properties of their elements, but in many cases new properties emerge, reflecting the new structural organization of their elements. This aspect of complex fluids has recently been investigated using colloidal particles in liquid crystals³, ferrosmectics⁴ and polymer-doped liquid crystals⁵. Here we focus on microemulsion, a typical complex fluid. Microemulsion is defined as a mixture of three essential components: two immiscible liquids and surfactant molecules. Typical examples are water-in-oil or oil-in-water emulsions. In the former, microdroplets of water (inverse micelles) are dispersed in oil, while in the latter microdroplets of oil are dispersed in water. The rheological properties of two liquids and the microstructure of phase-separating liquids (size and shape of domains and their spatial distribution) strongly affect the physical properties of the resulting microemulsions¹.

Here we investigate what happens if one of the fluid components is an anisotropic liquid such as nematic liquid crystal. We expect that the breakdown of the continuous rotational symmetry may induce interesting optical and mechanical properties. There are studies on nematic emulsions (of micrometre droplet size)⁶⁻⁹; however, to our knowledge there are no studies on 'nematic microemulsions' (of nanometre droplet size). This difference in droplet size of more than two orders of magnitude is expected to result in crucial differences between the emulsions.

Recently, the physical properties of spatially confined nematic liquid crystals¹⁰⁻¹² and nematic elastomers (ref. 13 and references therein, and ref. 14) have been intensively studied to reveal the elastic distortion effects induced by spatial confinement and random disorder on phase transitions of liquid crystals. We expect a nematic microemulsion to provide a new random disorder effect on nematic ordering, essentially different from nematic emulsions and microemulsions.

We study here a lyotropic inverse micelle phase composed of water, oil (thermotropic liquid crystal, pentylcyanobiphenyl, 5CB) and surfactant (didodecyl dimethyl ammonium bromide, DDAB). When 5CB is in the isotropic phase, this system is equivalent to conventional water-in-oil microemulsions. However, when nematic ordering occurs below an isotropic to nematic (I-N) phase transition temperature (T_{IN}), inverse micelles dispersed in the nematic medium disturb long-range nematic order. Thus, we expect that a new type of superstructure may be produced by the strong anchoring effects of inverse micelles on nematic directors.

The composition of a mixture can be characterized by the weight fraction of DDAB in the lyotropic liquid crystal (DDAB+water), ϕ_{am} , and the weight fraction of 5CB to the total sample weight, $1 - \phi$. Two characteristic lengths, the radius of an inverse micelle a and the intermicellar distance d , can then be controlled 'independently' by changing two concentrations, ϕ_{am} and ϕ , respectively, as $a = \delta(3 - 2\phi_{am})/\phi_{am}$ and $d = (4\pi/3\phi)^{1/3}a$, where δ is the length of a DDAB molecule. In our experiments, we fixed ϕ_{am} to be 85%, so that a is kept constant ($a \approx 1.9$ nm). By changing ϕ from 0.02 to 0.5, we can vary d from ~ 15 nm to the inverse micelle size ($\sim 2a$). For large ϕ , the average distance between inverse micelles is thus only a few times longer than the length of a 5CB molecule. It should be noted that for $\phi > 0.3$ there may be a shape transition of inverse micelles from spherical to rodlike (cylindrical) micelles. The results for $\phi > 0.3$ will be presented elsewhere.

To study the phase behaviour of liquid-crystal microemulsions, we performed differential scanning calorimetry (DSC) measurements. The results are shown in Fig. 1. For mixtures, there appear two distinct peaks, indicating the existence of two thermodynamic transitions. We observed the phase behaviour with polarizing microscopy. The results are shown in Fig. 2. We can clearly see phase separation accompanying the appearance of birefringent nematic droplets below the onset temperature of the second DSC peak (T_{PS}) (see Figs 2c and d). Thus, the second DSC peak is assigned

to this phase separation. The first peak is assigned to the I-N transition because it becomes the I-N transition peak in the limit of $\phi \rightarrow 0$. A mixture always looks isotropic in visible light above T_{PS} and there is no visible change across the I-N transition under polarizing microscopic observation (see Figs 2a and b); this strongly indicates that in the temperature region between the onset temperature of the first peak (T_{IN}) and that of the second one (T_{PS}), there exists a new phase, which is isotropic at a large length scale, but locally has nematic order. This optically isotropic nature may be realized because nematic fluctuations are pinned as in the case of nematic elastomers. We call this new phase the transparent nematic (TN) phase. A typical distance between inverse micelles is much shorter than visible light wavelengths, and inverse micelles do not aggregate in this phase, so that there is no true long-range nematic order in the usual sense. At the same time, however, the distinct DSC peak just below T_{IN} indicates that this phase has a nematic order, at least locally. Thus we conclude that the TN phase macroscopically exhibits isotropic nature, but locally possesses nematic order. Below T_{PS} , the system phase separates into nematic (N) and TN phases. The nematic droplets emerging from the TN phase are strongly birefringent (see Figs 2c and d), as expected.

To unambiguously establish that the TN phase is optically transparent, we also made direct visual observation of a macroscopic sample in a glass tube with a diameter of 10 mm, immersed in a temperature-controlled bath. The temperature was calibrated by using the I-N transition of pure 5CB. We confirm that the sample is optically transparent and does not have any turbidity even below T_{IN} as long as the temperature is above T_{PS} . At T_{PS} , the system becomes completely turbid upon cooling. This conclusion is supported below.

We made dynamic polarized light scattering measurements of a macroscopic sample whose dimensions are about 10 mm $\times$ 10 mm $\times$ 30 mm. We note that the system is completely homogeneous on the macroscopic scale and there is no noticeable change in the scattering intensity above T_{PS} . The temperature dependence of the scattered intensity for $\phi = 0.071$ is shown in Fig. 3. Thus, we re-confirm that the TN phase is optically transparent and completely homogeneous above T_{PS} . The transparency of the TN phase is comparable to that of the isotropic phase. The time correlation

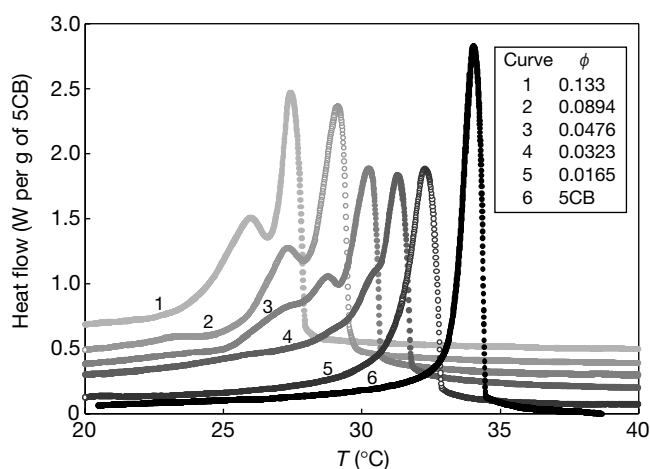


Figure 1 Concentration dependence of differential scanning calorimetry scans. The cooling rate was fixed at $0.3^\circ\text{C min}^{-1}$. Although only one peak is observed at T_{IN} for pure 5CB, two distinct peaks are detected for mixtures. The heat flow is scaled by the total amount (weight) of 5CB in a mixture. We take the onset of the increase of heat flow upon cooling as a transition temperature ($T_{IN}(\phi)$). The lower temperature transition ($T_{PS}(\phi)$) is assigned to phase separation by comparing it with the polarizing microscopic observation (see Fig. 2), while the higher temperature one is assigned to the isotropic–nematic transition as it is extrapolated to T_{IN} of pure 5CB in a dilute limit of the inverse micelle concentration.

function of composition fluctuations, $S(q, t)$ (where q is wave number) is well described by a single exponential function: $S(q, t) = S(q, 0)\exp(-2\Gamma t)$, where Γ is the decay rate of composition fluctuations. Thus, we obtain Γ by fitting an exponential function to the observed time correlation function. The temperature dependence of Γ for $\phi = 0.071$ is shown in Fig. 3. Γ decreases upon cooling till T_{IN} , but increases below T_{IN} . We found that this distinct minimum of Γ is located at T_{IN} , independently determined by DSC. This feature is also observed for $\phi = 0.063$. Thus, we suppose that the observed minimum of Γ reflects the phase transition between the two optically transparent phases, the isotropic phase and the TN phase. We speculate that the increase of Γ below T_{IN} is due to long-range repulsive interactions between micelles mediated by medium-range liquid-crystalline order. In relation to this, we briefly discuss the origin of the decay of composition fluctuations. The diffusion constant D estimated from the relation $\Gamma = Dq^2$ is of the order of $10^{-8} \text{ cm}^2 \text{ s}^{-1}$. This is of the same order as that determined from the relation $D = k_B T / 6\pi\eta a$, where k_B is Boltzmann's constant and η is the shear viscosity. This suggests that the decay is primarily due to the diffusional motion of inverse micelles in 5CB. However, we need to study the q dependence of Γ for a conclusive argument. In particular, we need such information to clarify the dynamics of inverse micelles in a transparent nematic phase below T_{IN} , which may not be described by simple diffusion. This point will be studied in the near future. Because dynamic light scattering experiments are

made in an equilibrium state (we wait about an hour at each temperature for equilibration), we can say that there exists a well defined thermodynamic transition between the isotropic and TN phases. That is, the TN phase is not a metastable phase (such as a supercooled isotropic phase), but a new thermodynamic phase with local nematic order. Below T_{PS} , the system becomes turbid because of phase separation. After several hours, it macroscopically phase-separates into two phases, the transparent TN phase (top) and the milky nematic phase (bottom), due to gravity.

To summarize all the experimental observations, a phase diagram is constructed as shown in Fig. 4. Here we also include the transition temperatures determined from the change in the temperature dependence of the intensity of static depolarized light scattering (data not shown). It demonstrates the ϕ dependence of $T_{IN}(\phi)$ and $T_{PS}(\phi)$, observed by various methods. The N–TN phase separation may be induced by competition between the elastic energy due to distortion and the entropy associated with the spatial configuration of inverse micelles. With decreasing T , the deformation costs more elastic energy, owing to further nematic ordering. Thus, a homogeneous transparent nematic phase becomes unstable and eventually phase-separates into an almost pure nematic phase and an isotropic TN phase filled with concentrated inverse micelles. The physical mechanism responsible for the new TN phase, or the I–TN transition, is not clear, although it may be true that the distortion of the director fields by inverse micelles is crucial to the appearance of the TN phase.

Now we propose an internal structure of this new phase, provided that liquid-crystal molecules are strongly anchored normal to the surface of an inverse micelle and that nematic directors are thus locally distorted by randomly dispersed inverse micelles. It is known that the relevant length of anchoring is given by $a_c \sim K/W$ (where W is molecular anchoring strength)¹⁵. Thus, W should be of the order of 10^{-3} N m^{-1} to realize strong anchoring in our system, provided that the characteristic length scale a_c is the order of 10 nm. This value corresponds to the upper limit of W reported so far^{16,17}. We note that W for the interface between water and 5CB in the presence

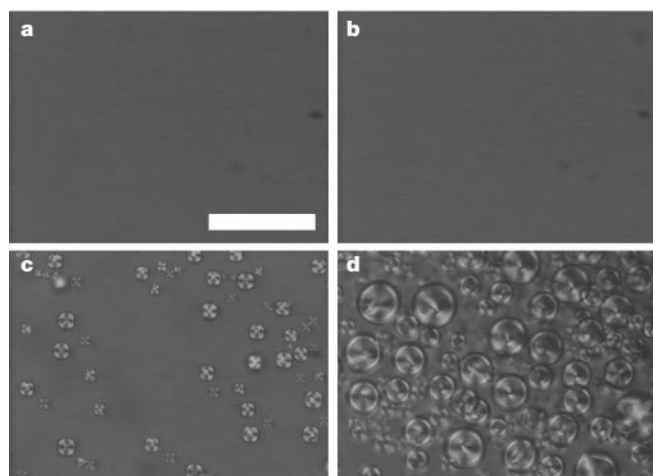


Figure 2 Polarizing microscope observation of a sample of $\phi = 0.08$ at various temperatures. **a**, $T = 32.0^\circ\text{C}$; **b**, $T = 29.5^\circ\text{C}$; **c**, $T = 28.0^\circ\text{C}$; and **d**, $T = 27.5^\circ\text{C}$. For this mixture, $T_{IN} = 30.0^\circ\text{C}$ and $T_{PS} = 28.2^\circ\text{C}$. The scale bar corresponds to $500 \mu\text{m}$. A sample was sandwiched between two cover glasses, whose gap was set to 1 mm . In this measurement, we used a thick sample to avoid the possible effects of homeotropic anchoring. This is because in a thin sample homeotropic alignment of liquid crystal may be induced, which might make a nematic state indistinguishable from an isotropic state under polarizing microscopy observation. We confirmed that pure 5CB indeed exhibits the texture characteristic of a nematic phase below T_{IN} in this sample geometry. For a mixture of $\phi = 0.08$, no structures were observed in the higher-temperature region, where 5CB is in the isotropic state (**a**). This clearly indicates that the 5CB molecules can be used as an isotropic oil component of microemulsions there. With decreasing temperature, the system should become a new phase below T_{IN} and above T_{PS} , according to our DSC measurements. Surprisingly, however, this new phase is optically indistinguishable from an isotropic state (see **b**), which strongly suggests that it is an optically transparent phase. With further decreasing temperature, phase separation takes place at $T_{PS}(\phi)$ and nematic droplets appear, as shown in **c**. This $T_{PS}(\phi)$ coincides well with the onset temperature of the second DSC peak. The volume fraction of the nematic phase increases with decreasing T (compare **c** and **d**), consistent with the phase diagram (see Fig. 4). Images **c** and **d** are taken during the coarsening process of phase separation. With increasing T , these nematic droplets disappear at the same temperature $T_{PS}(\phi)$, which implies that the transition is reversible.

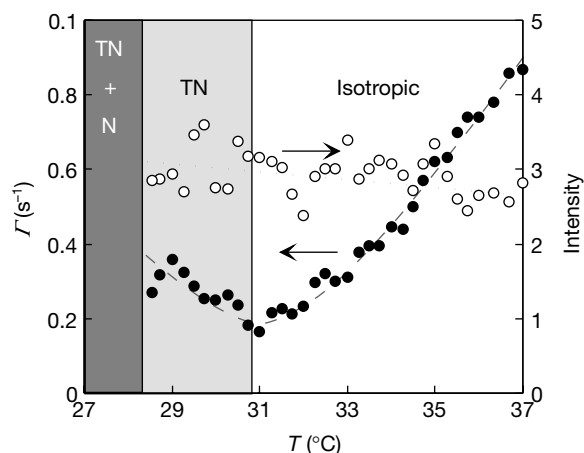


Figure 3 Results of dynamic polarized light scattering measurements. The figure shows the temperature dependence of the characteristic decay rate Γ (filled circles) of concentration fluctuations and the intensity of the scattered light (open circles) for a bulk sample of $\phi = 0.071$ at the scattering wave number of $q = 0.39 \mu\text{m}^{-1}$. The transition temperature T_{IN} (the phase boundary) is independently determined by DSC. The coincidence of the minimum of Γ with T_{IN} determined by DSC indicates that the transparent nematic (TN) phase is a thermodynamic equilibrium phase, and not a metastable phase. The increase in Γ with decreasing T below T_{IN} may be due to elastic interactions between inverse micelles induced by distortion of director fields. We stress that there is no drastic change in the scattering intensity across T_{IN} . This clearly demonstrates that the scattering intensity, or the transparency of the sample, of the TN phase is the same as that of an isotropic phase.

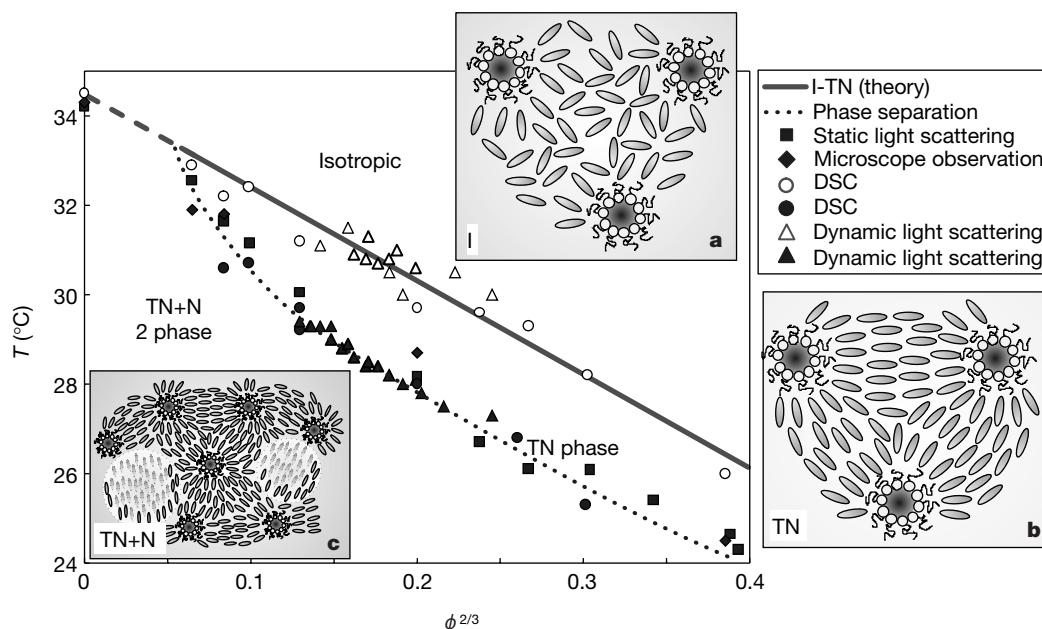


Figure 4 The $\phi - T$ phase diagram determined by experiment. Shown are experimental data from microscopic observation, static depolarized and dynamic polarized light scattering, and DSC measurements, together with the theoretical prediction (solid line) for $T_{IN}(\phi)$. IN and PS stand for isotropic–nematic phase transition and phase separation, respectively. The horizontal axis represents the characteristic quantity, $\phi^{2/3}$ (see text). A

dotted line represents the phase-separation boundary and is drawn as a guide to the eye. We also show schematic figures of the states of microemulsion above T_{IN} (a), below T_{IN} but above T_{PS} (b), and below T_{PS} (c). We assume that nematic liquid crystal is strongly anchored to inverse micelles (see text).

of the surfactant sodium dodecyl sulphate has been shown^{6,7,15} to be larger than 10^{-4} N m^{-1} . According to ref. 15, the control parameter Wa/K is proportional to Q^{-1} (where Q is the orientational order parameter) and increases as $Q \rightarrow 0$. It should be noted that Q may be small in the TN phase, as it is located near the I–N transition temperature. Thus, a strong-anchoring situation might be possible for our TN phase, although it seems rather too strong if we note that the inverse micelle itself is stabilized by the surface energy of the similar strength. If we focus on the random disorder effects of inverse micelles on nematic director fields through such strong anchoring, we can speculate that the I–N transition temperature decreases as a result of the disturbance to the nematic ordering induced by inverse micelles, as the system has to pay an extra energy cost associated with the distortion of nematic director fields around inverse micelles. The characteristic curvature associated with the deformation of director fields may be roughly the inverse of the intermicellar distance, which is estimated as $(3\phi/(4\pi a^3))^{1/3}$. Following ref. 18, the I–TN transition temperature, $T_{IN}(\phi)$, is expected to be shifted downward from T_{IN} of the pure system, $T_{IN}(0)$, by $(K/A)[3/(4\pi a^3)\phi]^{2/3}$, where K is the Frank elastic constant and A is a constant. As shown in Fig. 4, the agreement between the prediction that $T_{IN}(0) - T_{IN}(\phi) \propto \phi^{2/3}$ and the experimental results is satisfactory. However, the validity of this simple argument should be checked more carefully, as (1) the assumption of the strong anchoring might not be the case, as mentioned above, and (2) the applicability of the mean-field coarse-grained model to our system, which is characterized by the mesoscopic length scale, is not obvious. Here we point out another possible scenario, which focuses on topological ordering effects instead of disorder effects. A ‘topologically ordered’ isotropic phase has been predicted to form^{19,20} between the conventional isotropic and nematic phase at large suppression of topological defects and weak nematic interactions. Our new phase may be connected to this ‘topologically ordered’ isotropic phase. Further experimental and theoretical studies are necessary to elucidate the physical mechanism responsible for the appearance of the TN phase.

We hope that this study will provide a new way of constructing

useful exotic complex fluids. We speculate that external shear deformation or magnetic fields induce global nematic order, or the breakdown of the rotational symmetry in a length scale longer than light wavelengths. Such a field-induced transition has recently been predicted theoretically for frustrated systems such as nematic elastomers and spin glass¹⁴. □

Received 12 October; accepted 23 November 2000.

- Daoud, M. & Williams, C. E. (eds) *Soft Matter Physics* (Springer, Berlin, 1999).
- de Gennes, P. G. & Prost, J. *Physics of Liquid Crystals* 2nd edn (Clarendon, Oxford, 1993).
- Poulin, P., Raghunathan, V. A., Richetti, P. & Roux, D. On the dispersion of latex particles in a nematic solution. I. Experimental evidence and simple model. *J. Phys. II* **4**, 1557–1569 (1994).
- Fabre, P., Casagrande, C., Veyssie, M., Cabuil, V. & Massart, R. Ferromagnetic: A new magnetic and mesomorphic phase. *Phys. Rev. Lett.* **64**, 539–542 (1990).
- Ott, A., Urbach, W., Langevin, D., Ober, R. & Waks, M. Light scattering study of surfactant multilayers elasticity. Role of incorporated proteins. *Europhys. Lett.* **12**, 395–400 (1990).
- Lubensky, T. C., Pettry, D., Currier, N. & Stark, H. Topological defects and interactions in nematic emulsions. *Phys. Rev. B* **57**, 610–625 (1998).
- Poulin, P. & Weitz, D. A. Inverted and multiple nematic emulsions. *Phys. Rev. E* **57**, 626–637 (1998).
- Terentjev, E. M. Stability of liquid crystalline macroemulsions. *Europhys. Lett.* **32**, 607–612 (1995).
- Terentjev, E. M. in *Modern Aspects of Colloidal Dispersions* (eds Ottewill, R. H. & Rennie, A. R.) 257–267 (Kluwer Academic, London, 1998).
- Bellini, T. *et al.* Phase behavior of the liquid crystal 8CB in a silica aerogel. *Phys. Rev. Lett.* **69**, 788–791 (1992).
- Wu, L., Zhou, B., Garland, C. W., Bellini, T. & Schaefer, D. W. Heat-capacity study of nematic–isotropic and nematic–smectic-A transitions for octylcyanobiphenyl in silica aerogels. *Phys. Rev. E* **51**, 2157–2165 (1995).
- Copic, M. & Mertelj, A. Reorientation in random potential: A model for glasslike dynamics in confined liquid crystals. *Phys. Rev. Lett.* **80**, 1449–1452 (1998).
- Clarke, S. M., Terentjev, E. M., Kundler, I. & Finkelmann, H. Texture evolution during the polydomain–monodomain transition in nematic elastomers. *Macromolecules* **31**, 4862–4872 (1998).
- Fridrikh, S. V. & Terentjev, E. M. Order-disorder transition in an external field in random ferromagnets and nematic elastomers. *Phys. Rev. Lett.* **79**, 4661–4664 (1997).
- Kuksenok, O. V., Ruhwandl, R. W., Shiyankovskii, S. V. & Terentjev, E. M. Director structure around a colloid particle suspended in a nematic liquid crystal. *Phys. Rev. E* **54**, 5198–5203 (1996).
- Grawford, G. P., Ondris-Crawford, R. J., Doane, J. W. & Zumer, S. Systematic study of orientational wetting and anchoring at a liquid-crystal-surfactant interface. *Phys. Rev. E* **53**, 3647–3661 (1996).
- Seo, D.-S., Kobayashi, S., Kang, D.-Y. & Yokoyama, H. Effects of rubbing and temperature dependence of polar anchoring strength of homogeneously aligned nematic liquid crystal on polyimide Langmuir–Blodgett orientation films. *Jpn J. Appl. Phys.* **34**, 3607–3611 (1995).
- Iannacchione, G. S., Crawford, G. P., Zumer, S., Doane, J. W. & Finotello, D. Randomly constrained orientational order in porous glass. *Phys. Rev. Lett.* **71**, 2595–2598 (1993).

19. Lammert, P. E., Rokhsar, D. S. & Toner, J. Topology and nematic ordering. I. A gauge theory. *Phys. Rev. E* **52**, 1778–1800 (1995).
20. Toner, J., Lammert, P. E. & Rokhsar, D. S. Topology and nematic ordering. II. Observable critical behavior. *Phys. Rev. E* **52**, 1801–1810 (1995).

Acknowledgements

We thank M. Takahashi for DSC measurements. H.T. is also grateful to T. C. Lubensky, E. M. Terentjev and S. Zumer for valuable discussions.

Correspondence and requests for materials should be addressed to H.T. (e-mail: tanaka@iis.u-tokyo.ac.jp).

Increased thermohaline stratification as a possible cause for an ocean anoxic event in the Cretaceous period

Jochen Erbacher*, Brian T. Huber†, Richard D. Norris‡ & Molly Markey†§

* Bundesanstalt für Geowissenschaften und Rohstoffe, Referat Meeresgeologie, Stilleweg 2, 30655 Hannover, Germany

† Department of Paleobiology, NHB-121, National Museum of Natural History, Smithsonian Institution, Washington DC 20560, USA

‡ Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543-1541, USA

Ocean anoxic events were periods of high carbon burial that led to drawdown of atmospheric carbon dioxide, lowering of bottom-water oxygen concentrations and, in many cases, significant biological extinction^{1–5}. Most ocean anoxic events are thought to be caused by high productivity and export of carbon from surface waters which is then preserved in organic-rich sediments, known as black shales. But the factors that triggered some of these events remain uncertain. Here we present stable isotope data from a mid-Cretaceous ocean anoxic event that occurred 112 Myr ago, and that point to increased thermohaline stratification as the probable cause. Ocean anoxic event 1b is associated with an increase in surface-water temperatures and runoff that led to decreased bottom-water formation and elevated carbon burial in the restricted basins of the western Tethys and North Atlantic. This event is in many ways similar to that which led to the more recent Plio-Pleistocene Mediterranean sapropels, but the greater geographical extent and longer duration (~46 kyr) of ocean anoxic event 1b suggest that processes leading to such ocean anoxic events in the North Atlantic and western Tethys were able to act over a much larger region, and sequester far more carbon, than any of the Quaternary sapropels.

The mid-Cretaceous is associated with ocean anoxic events (OAEs)—periods of elevated carbon burial in marine sediments known primarily from Tethyan basins, the Atlantic and equatorial Pacific. These events have been interpreted as the result of increased surface-water productivity caused by either nutrient leaching of flooded lowlands during transgressions and/or increased continental runoff due to an accelerated hydrological cycle^{2,4,6,7}. Recent high-resolution stable isotope studies suggest an intensified vertical mixing and warming of the water column as a cause for the observed productivity increase during early Aptian OAE 1a and late Albian OAE 1d^{5,8,9}. For OAE 1d, this intensified mixing is expressed by decreased vertical isotopic gradients and a warming of the entire water-column. High-resolution stable isotope studies on planktic

and benthic foraminifers¹⁰ indicate that a bathyal warming and vertical mixing may have caused OAE 2 and its associated extinctions. Or some of the OAEs may reflect decreased vertical mixing and the development of euxinic conditions throughout the water column that permits organic carbon to accumulate on the sea floor^{11,12}.

Ocean Drilling Program (ODP) Leg 171 B drilled mid-Cretaceous black shales of Albian age in the western subtropical Atlantic, off Florida. ODP Hole 1049C recovered Aptian to Eocene sediments, including a 46-cm-thick succession of laminated black shale, representing early Albian OAE 1b (ref. 13). Total organic carbon (TOC) contents in laminated black shales range from 0.5 to 12.3% (ref. 14). The record of OAE 1b at Site 1049 is unusual for most OAE sediments because the foraminifera are extremely well preserved and can be used to study the geochemical record of the event. Both planktic and benthic species have glassy shells with preserved surface ornamentation and without infilling calcite. In addition, the stable isotope signature of individual species suggests that stable isotopic ratios were not homogenized, which would be expected for diagenetically altered calcite.

Oxygen isotope values for the planktic species *Hedbergella aff. Hedbergella trocoidea* are consistently more negative by 0.7 to 0.1‰ and carbon isotope values heavier by 0.60 to 0.05‰ than for the planktic species *Hedbergella speetonensis tuniensis*, suggesting a deeper habitat for the latter (see Supplementary Information). Benthic species show a constant offset of carbon isotopes with heavier values for *Osangularia schloenbachi* and negative values for *Gyroidinoides nitidus*. This suggests an infaunal habitat for *G. nitidus* and an epifaunal mode of life for *O. schloenbachi*, as infaunal forms tend to yield more negative signals owing to the enrichment of the pore waters with isotopically negative carbon that is derived from oxygenized organic carbon.

We identified four intervals (events I to IV, Fig. 1) that represent significant changes in ocean temperature, salinity and carbon flux during the OAE.

Event I is characterized by stable carbon and oxygen isotope values. Oxygen isotope values are surprisingly positive, suggesting either very cool surface and deep-water temperatures (around 12 °C for surface water and 9 °C for bottom water of an upper bathyal water depth around ~1,000 m) and/or high salinities before the black shale. In addition, oxygen and carbon isotope gradients between the surface and bottom water are very small, indicating a uniform water mass in the upper 1,000 m of the water column during this phase.

Event II marks a large 1.6‰ negative oxygen isotopic shift in planktic foraminifers below the base of the black shale. This decrease is paralleled by a 0.9‰ increase of planktic $\delta^{13}\text{C}$ values, followed by a decrease of the same magnitude at the base of the black shale. Benthic values in this interval remain relatively constant for the carbon isotopes, but oxygen isotopic values of benthic foraminifers show a slight decrease in the lowermost part of the black shale that occurs later than the negative shift of planktic foraminifers. The steep decrease in oxygen isotope values below the black shale is difficult to explain with a rise of sea surface temperatures only, as it would reflect a geologically rapid and significant rise of approximately 10 °C. However, this decrease could be explained by a combination of salinity decrease and temperature increase of surface waters. The small increase in the vertical carbon gradient may point to an increase of surface-water productivity before the black shale, leading to increased carbon export and greater carbon isotope gradients between surface and bottom water. This interpretation is supported by the synchronous presence of benthic foraminifers that indicate high organic carbon fluxes to the sea floor^{15,16}. The decrease of carbon isotope values at the base of the black shale is difficult to explain, but might be connected to the presence of negative organic carbon in the black shale. However, constant sedimentation rates and the lack of compositional changes

§ Present address: Department of Invertebrate Paleontology, Museum of Comparative Zoology, 26 Oxford Street, Harvard University, Cambridge, Massachusetts 02138, USA.

of organic matter between marls and black shales suggest that productivity was not a primary factor controlling the positive carbon excursion¹⁴. An alternative cause for the positive carbon shift is degassing of carbon dioxide from surface water, coupled with a decreased supply of isotopically negative carbon from deep water during the observed warming of surface waters (M. A. Arthur, personal communication).

Event III marks a brief increase of planktic oxygen isotope values while benthic oxygen isotopes and carbon isotope values for both benthic and planktic foraminifers remain stable. This rise of $\delta^{18}\text{O}$ values may reflect a short drop in temperature and/or a rise in surface-water salinity and is paralleled by a repopulation of benthic foraminifers indicating a synchronous reoxygenation at the sea floor^{15,16}.

Event IV marks the long-term evolution towards pre-event values of planktic oxygen isotope values in the upper part and above the black shale. We note that the re-establishment of low oxygen isotope gradients from the surface to deep water does not coincide with the top of the black shale, but occurs about 1.4 m above it. This discrepancy might be explained by "burn down" of organic matter following reoxygenation of the sea floor, as was observed in modern Mediterranean sapropels¹⁷. However, reoccurrence of benthic foraminifers just above the black shale suggests termination of bottom-water dysoxia at the top of the preserved black shale.

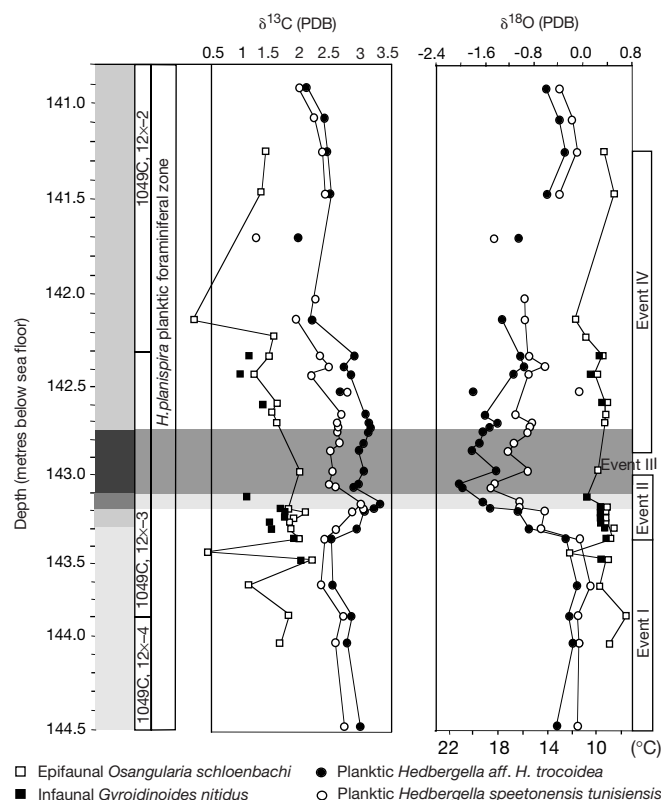


Figure 1 Lithology, and oxygen and carbon isotopes of planktic and benthic foraminifers plotted against depth of ODP Hole 1049C, western subtropical Atlantic. The benthic foraminiferal isotope record is shown in open (epifaunal *Osangularia schloenbachii*) and filled squares (infaunal *Gyroidinoides nitidus*), and the planktic foraminiferal isotope record is shown in open (*Hedbergella speetonensis tunisiensis*) and filled circles (*Hedbergella aff. H. trochoidea*). The shaded bar indicates the laminated black shale interval representing OAE 1b (dark: distinctly laminated; light: not laminated). Palaeotemperatures were calculated using the equation of Erez and Luz²⁷ and the standard mean ocean water (SMOW)/PDB conversion of -0.27 (ref. 28), assuming $-1.2\text{‰}_{\text{SMOW}}$ for a non-glacial world²⁹. Four events (I to IV) are identified in the OAE 1b stable isotope record. Nomenclature: '1049C, 12 × -4' indicates ODP Hole 1049C, Core 12 ×, Section 4.

Our observations suggest that a well mixed, well ventilated, low-nutrient water column with relatively high salinities and cool temperatures increased moderately in temperature at the same time as surface-water salinity decreased, causing pronounced thermohaline stratification of the water column. This is documented by a large increase in the oxygen isotope gradients from the surface to deep water observed in our record and suggests that black shale deposition was triggered by a reduction in ventilation of subthermocline water. The termination of OAE 1b was caused by a gradual reduction of vertical temperature and salinity gradients. The increased vertical oxygen isotopic gradient cannot represent an evolutionary change in depth habitats of the analysed planktic organisms, as, first, a synchronous depth-habitat modification of two different species is very unlikely and, second, the observed change is reversed above the dark shales.

Tuned cycle stratigraphy yielded¹⁸ calculated sedimentation rates of approximately 1.0 cm kyr^{-1} for the early Albian interval of ODP Site 1049. According to these data the drop of $\delta^{18}\text{O}$ values during event II lasted approximately 32 kyr, the rise of $\delta^{18}\text{O}$ values (event IV) persisted for 175 kyr, and the duration of black shale deposition was about 46 kyr. The total duration of the OAE 1b interval (from base of event II to the top of event IV) is estimated to be approximately 210 kyr.

It has been suggested that OAE 1b was caused by an increase in surface-water productivity similar to other OAEs^{15,19}. In contrast to other OAEs, which have mostly been explained by an intensified vertical mixing of intermediate and surface waters, our data show that productivity seems to play a minor role as a factor controlling OAE 1b.

The importance of variation in precipitation and/or continental runoff as a mechanism for changing deep water convection has been demonstrated in a series of ocean model experiments²⁰. Such hydrologic changes have been attributed as the cause for formation of sapropels in the Mediterranean during the Quaternary. We consider this model of sapropel formation to be the best analogue to black shale formation during OAE 1b.

In the modern and Quaternary Mediterranean, vertical temperature gradients are very low, as can be observed during our event I. During sapropel formation, a steep decrease of $\delta^{18}\text{O}$ values observed in planktic foraminifera has been related to a drop of surface-water salinity and minor increases in surface-water temperatures²¹. The increased surface stratification is believed to be caused by increased continental runoff during warm periods, resulting in a positive freshwater balance and a switch from an anti-estuarine to estuarine circulation in the Mediterranean^{21,22}. Such an estuarine circulation system is controlled by the presence of sills in the gateways between the western and eastern Mediterranean Sea and the Atlantic Ocean.

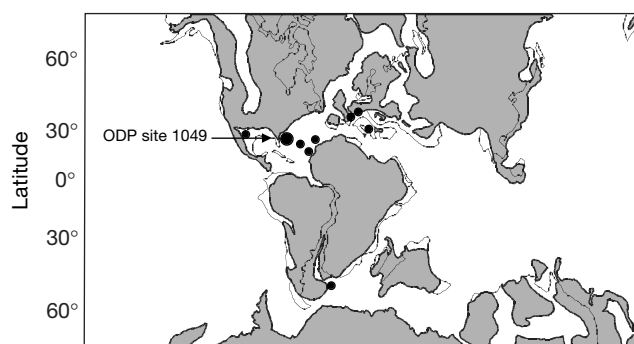


Figure 2 Palaeogeographic reconstruction for the late Aptian (120 Myr ago) with the positions of the biostratigraphically well dated, known OAE 1b occurrences. Modified after refs 19 and 30. We note that OAE 1b sections are only known from Tethyan basins, North and South Atlantic.

Recent investigations documented reoxygenation horizons from several Mediterranean sapropels, similar to the one we described, which are considered to be the result of short (approximately 500 years) phases of vertical mixing. Palaeotemperature records document a phase of cool surface-water temperatures during these interruptions, again similar to what our data suggest for the reoxygenation during event III of OAE 1b (ref. 25).

Mediterranean sapropels are the result of monsoonal fluctuation during the Quaternary resulting in a cyclic presence of numerous sapropels. In contrast to Site 1049, where the described black shale horizon is the only carbonaceous layer, late Aptian to early Albian sections in the western Tethyan basins show numerous black shales that have been described as being cyclic^{24,25}. A comparable situation can be observed in the Quaternary Mediterranean, where only the best developed sapropels in the restricted eastern basin are present in the more open western basin²⁶.

We believe that the mechanisms leading to OAE 1b reflect conditions similar to those leading to Quaternary Mediterranean sapropels. However, the geographical extent of OAE 1b was much larger and, according to the age model available at present, the duration of OAE 1b was much longer, suggesting the North Atlantic/western Tethyan Ocean remained in a "super-sapropel stage" for approximately four times longer than the longest of the Plio-Pleistocene Mediterranean sapropels. The great extent of OAE 1b (Fig. 2) suggests that processes that restricted ventilation of the North Atlantic and Tethys were able to act over a much larger region than ever occurred during the Plio-Pleistocene interval of sapropel formation. We suggest that the partial tectonic isolation of the various basins in the Tethys and Atlantic, a low sea level and the initiation of warm global climates may be important factors in setting up oceanic stagnation during OAE 1b. □

Received 10 July; accepted 8 November 2000.

- Arthur, M. A., Brumsack, H.-H., Jenkyns, H. C. & Schlanger, S. O. in *Cretaceous Resources, Events and Rhythms* (eds Ginsburg, R. N. & Beaudoin, B.) 75–119 (Kluwer, Dordrecht, 1990).
- Erbacher, J., Thirum, J. & Littke, R. Evolution patterns of radiolaria and organic matter variations—a new approach to identify sea-level changes in mid-Cretaceous pelagic environments. *Geology* **24**, 499–502 (1996).
- Kuypers, M. M. M., Pancost, R. D. & Sinninghe Damsté, J. S. A large and abrupt fall in atmospheric CO₂ concentrations during Cretaceous times. *Nature* **399**, 342–345 (1999).
- Jenkyns, H. C. Mesozoic anoxic events and paleoclimate. *Zbl. Geol. Paläont. Teil 1*, 7–9, 943–949 (1997).
- Hochuli, P. A. et al. Episodes of high productivity and cooling in the early Aptian Alpine Tethys. *Geology* **27**, 657–666 (1999).
- Thirum, J., Brumsack, H.-H., Rullkötter, J. & Meyers, P. The Cenomanian/Turonian Boundary Event in the Indian Ocean—a key to understand the global picture. *Geophys. Monogr.* **70**, 253–273 (1992).
- Weissert, H., Lini, A., Föllmi, K. B. & Kuhn, O. Correlation of Early Cretaceous carbon isotope stratigraphy and platform drowning events: a possible link? *Paleogeogr. Palaeoclim. Palaeoecol.* **137**, 189–203 (1998).
- Wilson, P. A., Norris, R. D. & Erbacher, J. Tropical sea surface temperature records and black shale deposition in the mid-Cretaceous western Atlantic (Blake Nose and Demerara Rise). *Eos* **80**, F488 (1999).
- Menegatti, A. P. et al. High-resolution $\delta^{13}\text{C}$ stratigraphy through the early Aptian "Livello Selli" of the Alpine Tethys. *Paleoceanography* **13**, 530–545 (1998).
- Huber, B. T., Leckie, R. M., Norris, R. D., Bralower, T. J. & CoBabe, E. Foraminiferal assemblages and stable isotopic change across the Cenomanian-Turonian Boundary in the subtropical North Atlantic. *J. Foram. Res.* **29**, 392–417 (1999).
- Sinninghe Damsté, J. S. & Köster, J. A euxinic southern North Atlantic Ocean during the Cenomanian/Turonian oceanic anoxic event. *Earth Planet. Sci. Lett.* **158**, 165–173 (1998).
- Fischer, A. G. & Arthur, M. A. Secular variations in the pelagic realm. *SEPM Spec. Publ.* **25**, 19–50 (1977).
- Norris, R. D. et al. Blake Nose paleoceanographic transect, Western North Atlantic. *Proc. ODP Init. Rep. B* **171**, 1–749 (1998).
- Barker, C. E., Pawlewicz, M. & Cobabe, E. A. in *Western North Atlantic Paleogene and Cretaceous Paleooceanography* (eds Kroon, D., Norris, R. D. & Klaus, A.) (Geological Society, London, in the press).
- Erbacher, J., Hemleben, C., Huber, B. & Markey, M. Correlating environmental changes during early Albian oceanic anoxic event 1b using benthic foraminiferal paleoecology. *Mar. Micropal.* **38**, 7–28 (1999).
- Holbourn, A., Kuhnt, W. & Erbacher, J. Benthic foraminifera from lower Albian black shales (Site 1049, ODP Leg 171): evidence for a non "uniformitarian" record. *J. Foram. Res.* (in the press).
- Mercone, D. et al. Duration of S1, the most recent sapropel in the eastern Mediterranean Sea, as indicated by accelerator mass spectrometry radiocarbon and geochemical evidence. *Paleoceanography* **15**, 336–347 (2000).

- Ogg, J. G., Röhl, U. & Geib, T. Astronomical tuning of Aptian-Albian boundary interval: oceanic anoxic event 1b through lower Albian magnetic suchron M²-2^r. *Eos* **80**, F491–F492 (1999).
- Bralower, T. J. et al. Dysoxic/Anoxic episodes in the Aptian-Albian (Early Cretaceous). *Geophys. Monogr.* **77**, 5–37 (1993).
- Bice, K. L., Barron, E. J. & Peterson, W. H. Continental runoff and early Cenozoic bottom-water sources. *Geology* **25**, 951–954 (1997).
- Emeis, K.-C. et al. Stable isotope and alkenone temperature records of sapropels from Sites 964 and 967: constraining the physical environment of sapropel formation in the eastern Mediterranean Sea. *Proc. ODP Sci. Res.* **160**, 309–331 (1998).
- Rohling, E. J. Review and new aspects concerning the formation of eastern Mediterranean sapropels. *Mar. Geol.* **122**, 1–28 (1994).
- Myers, G. P. & Rohling, E. J. Modeling a 200-yr interruption of the Holocene sapropel S1. *Quat. Res.* **53**, 98–104 (2000).
- Wortmann, U. G., Hesse, R. & Zacher, W. Major-element analysis of cyclic black shales: Paleoenvironmental implications for the Early Cretaceous deep western Tethys. *Paleoceanography* **14**, 525–541 (1999).
- De Boer, P. L. Aspects of middle cretaceous pelagic sedimentation in Southern Europe. *Geol. Ultrajectina* **31**, 1–112 (1983).
- Meyers, P. A. & Doose, H. Sources, preservation and thermal maturity of organic matter in Pliocene-Pleistocene organic-carbon-rich sediments of the western Mediterranean Sea. *Proc. ODP Sci. Res.* **161**, 383–390 (1999).
- Erez, J. & Luz, B. Experimental paleotemperature equation for planktonic foraminifera. *Geochim. Cosmochim. Acta* **47**, 1025–1031 (1983).
- Hut, G. Consultants group meeting on stable isotope reference samples for geochemical and hydrological investigations. 1–42 (Report to Director General of the Institute of Atomic Energy Agency, Vienna, 1987).
- Shackleton, N. J. & Kennett, J. P. Paleotemperature history of the Cenozoic and the initiation of Antarctic glaciation: oxygen and carbon isotope analyses in DSDP Sites 277, 279 and 281. *DSDP Init. Rep.* **29**, 743–755 (1975).
- Barron, E. J., Harrison, C. G., Sloan, J. L. & Hay, W. W. Paleogeography, 80 million years ago to present. *Ecol. Geol. Helv.* **74**, 443–470 (1981).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to J.E. (e-mail: erbacher@bgr.de).

300-Myr-old magmatic CO₂ in natural gas reservoirs of the west Texas Permian basin

Chris J. Ballentine*, Martin Schoell†, Dennis Coleman‡ & Bruce A. Cain§

* IGMR, Dept Erdwissenschaften, ETH Zentrum NO CO 61.7, 8092 Zürich, Switzerland

† Chevron Research and Technology Company, 6001 Bollinger Canyon, San Ramon, California 94583, USA

‡ Isotech Laboratories Inc., 1308 Parkland Court, Champaign, Illinois 61821-1826, USA

§ Altura Energy LLP, Houston, Texas 77210-4294, USA

Except in regions of recent crustal extension¹, the dominant origin of carbon dioxide in fluids in sedimentary basins has been assumed to be from crustal organic matter² or mineral reactions^{3,4}. Here we show, by contrast, that Rayleigh fractionation caused by partial degassing of a magma body can explain the CO₂/He ratios and $\delta^{13}\text{C}(\text{CO}_2)$ values observed in CO₂-rich natural gases in the west Texas Val Verde basin and also the mantle ³He/²²Ne ratios observed in other basin systems⁵. Regional changes in CO₂/He and CO₂/CH₄ ratios can be explained if the CO₂ input pre-dates methane generation in the basin, which occurred about 280 Myr ago⁶. Uplift to the north of the Val Verde basin between 310 and 280 Myr ago⁶ appears to be the only tectonic event with appropriate timing and location to be the source of the magmatic CO₂. Our identification of magmatic CO₂ in a foreland basin indicates that the origin of CO₂ in other mid-continent basin systems should be re-evaluated. Also, the inferred closed-system preservation of natural gas in a trapping structure

for ~300 Myr is far longer than the residence time predicted by diffusion models^{7,8}.

Identification of magmatic CO₂ and the processes responsible for its input into near-surface reservoirs are critical in providing both a sample resource for the investigation of mantle volatile evolution⁹ and in helping to develop regional strategies for natural gas exploration. Discriminating between magmatic and crustal sources of CO₂ is not possible with stable isotopes alone, because the ranges of $\delta^{13}\text{C}$ in magmatic and crustal CO₂ are non-unique^{1,10}. In contrast, ³He is an unambiguous tracer of magmatic volatiles¹¹. Because the rate of He diffusion is relatively low in basalt melt or carbon–oxygen–hydrogen fluids¹², magmatic ³He must be carried, or advected, into shallow systems. The carrier fluid is inferred to be CO₂ (see, for example, ref. 1). The CO₂/³He ratio of the mantle is $(1-2) \times 10^9$, while the CO₂/³He range found in magmatic samples, $(1-10) \times 10^9$, probably reflects variable degrees of degassing and different stages of magma evolution¹³⁻¹⁶. Nevertheless, this range is small compared to those observed in natural gases—between 10^5 and 10^{13} (ref. 1).

Although the CO₂ in natural gases that preserve CO₂/³He ratios in the range $(1-10) \times 10^9$ is probably dominated by magmatic gas, the possibility of a subsequent crustal CO₂ overprint or CO₂ loss in these gases must also be considered. Systematic behaviour of CO₂/³He, ³He/⁴He and stable isotopes allow us to evaluate these possibilities. Classic CO₂ occurrences in natural gases in the mid-continent USA provide an ideal location in which the importance of magmatic CO₂ input into basin systems can be investigated. The west Texas Permian basin was formed as a result of the late Palaeozoic collision of South America with North America that also resulted in the Marathon–Ouachita orogenic belt as well as widespread interior continental deformation. The Val Verde basin is a foreland sub-basin of the west Texas Permian basin, and lies between the Central basin platform and the Marathon thrust belt

(MTB)¹⁷⁻¹⁹ (Fig. 1). Natural gas reservoirs show a systematic regional increase in CO₂ content towards the MTB, varying from an average of about 3% in the basin centre to as high as 97% on the foredeep margin of the thrust belt. The main producing formation in the JM Brown Bassett (JM-BB) field is brecciated Ordovician Ellenberger dolomite. This field reflects the regional spatial trend in natural gas CO₂ content, with samples increasing from 20% to 55% CO₂ towards the MTB. The remaining gas is dominated by CH₄. We present a study of C, He, Ne and Ar isotopes of representative producing wells across this field (Table 1).

The ³He/⁴He ratio correlates directly with CO₂ content, and shows two-component mixing between CH₄ and CO₂ endmembers (Fig. 2). CO₂/³He varies between 4.2×10^9 and 6.2×10^9 , is within the magmatic range, and is also correlated with CO₂ content (Fig. 3). But the systematic variation in CO₂/³He is not predicted by simple mixing, and must be accounted for. The crustal-radiogenic noble gases associated with the CO₂ endmember would be expected to reflect transport or phase-related fractionation of the CO₂/³He values. Measured ²⁰Ne/²²Ne, ²¹Ne/²²Ne and ⁴⁰Ar/³⁶Ar enable crustal-radiogenic ²¹Ne and ⁴⁰Ar to be resolved from air-derived components²⁰. Crustal-radiogenic ⁴He can be resolved from the ³He/⁴He ratio (Fig. 2). Average crustal-radiogenic ⁴He/⁴⁰Ar and ⁴He/²¹Ne values for the JM-BB field are 4.63 ± 0.57 and $1.76 \pm 0.10 \times 10^7$, respectively, and are indistinguishable from ⁴He/⁴⁰Ar calculated assuming average crust concentrations of U, Th and K (⁴He/⁴⁰Ar = 4.9; ref. 20), and observed average crust ⁴He/²¹Ne values (⁴He/²¹Ne = 1.7×10^7 ; ref. 21). There is no correlation between CO₂/³He variation and crustal radiogenic gases and we can, therefore, rule out transport or phase-related fractionation. Three mechanisms potentially explain the variation in CO₂/³He: (1) crustal CO₂ addition; (2) precipitation of CO₂; or (3) magmatic source variation.

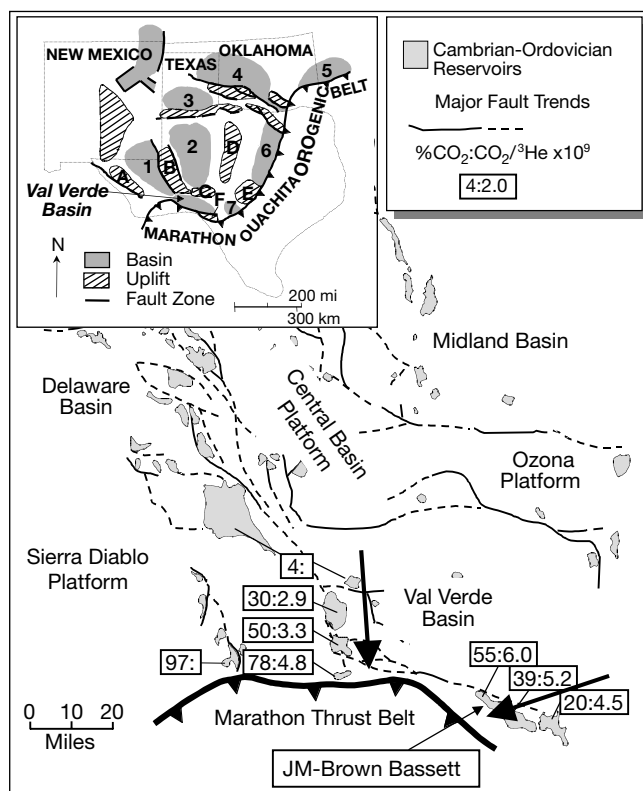


Figure 1 Location of the JM-Brown Bassett (JM-BB) natural gas field. Arrows show the direction of the regional increase in CO₂ content and CO₂/³He ratio towards the Marathon thrust belt (MTB). CO₂ and ³He data for the JM-BB field are from this study. Other data are from Chevron (unpublished). Figure after Shumaker¹⁷. Inset, location of the Val Verde

basin relative to the major Permian basin and uplift features after Montgomery¹⁹. Basins: 1, Delaware; 2, Midland; 3, Palo-Duro; 4, Anadarko; 5, Arkoma; 6, Ft Worth; 7, Kerr. Uplifts: A, Sierra Diablo; B, Central basin; C, Ozona; D, Concho arch; E, Llano; F, Devils River.

We show in Fig. 3 the effects that CO₂ addition, CO₂ loss and CH₄ addition (simple mixing) would have on the data distribution. None of these processes alone can account for the systematic change in CO₂/He as a function of CO₂ content. We can use the $\delta^{13}\text{C}(\text{CO}_2)$ values to investigate further the effect of crustal CO₂ addition or precipitation reactions. $\delta^{13}\text{C}(\text{CO}_2)$ and $\delta^{13}\text{C}(\text{CH}_4)$ show very little variation across the JM-BB field, with values varying between -2.70 and -3.06‰ and -37.26 and -37.68‰, respectively (Table 1). Inorganic isotopic equilibrium between CO₂ and CH₄ does not occur on geological timescales at temperatures less than 200 °C (ref. 22). The JM-BB field is at ~100 °C, suggesting that the CO₂ and CH₄ stable isotopes reflect the composition of the fluids input into the reservoir. 50% crustal CO₂ would have to be added to the system to change the CO₂/He from 4×10^9 to 6×10^9 . This added CO₂ would have to have a $\delta^{13}\text{C}$ value to within $\pm 0.2\text{‰}$ of the magmatic component to preserve the observed $\delta^{13}\text{C}(\text{CO}_2)$ range. Similarly, precipitation of CO₂ to reduce the CO₂/He from 6×10^9 to 4×10^9

would also result in isotopic fractionation. Although $\delta^{13}\text{C}(\text{CO}_2)$ fractionation between CO₂ and calcite by equilibration via an aqueous phase is strongly temperature dependent, we can take a typical value of 2.7‰. A gas phase precipitating 30% of its mass as calcite would change in $\delta^{13}\text{C}(\text{CO}_2)$ by 0.7‰ or 0.95‰ for equilibration or Rayleigh fractionation, respectively. We do not see this variation in the data set. We cannot rule out the possibility that crustal CO₂ with just the right $\delta^{13}\text{C}(\text{CO}_2)$ has been added to the system, nor that precipitation under just the right conditions has occurred to effect no isotopic fractionation. Nevertheless, combined with the complex process required to account for the systematic CO₂/He change with CO₂ content (Fig. 3), we consider the processes of crustal CO₂ addition or magmatic CO₂ loss to be very unlikely.

Rayleigh fractionation caused by partial degassing of a deep-

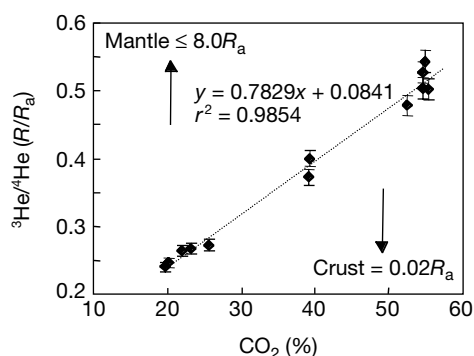


Figure 2 Plot of $^3\text{He}/^4\text{He}$ against CO₂ content. This figure shows two-component mixing between CH₄ and CO₂ endmembers with $^3\text{He}/^4\text{He}$ values of $0.08R_a$ and $0.85R_a$, respectively, where R is the measured $^3\text{He}/^4\text{He}$ and R_a is the $^3\text{He}/^4\text{He}$ ratio of air (1.4×10^{-6}). Taking crustal and mantle endmembers to be $0.02R_a$ and $8.0R_a$, respectively, 0.8% of the ^4He and 76% of the ^3He in the CH₄ endmember is magmatic. The higher $^3\text{He}/^4\text{He}$ in the CO₂ endmember show that in this component 10.6% of the ^4He , and 98.2% of the ^3He , is magmatic.

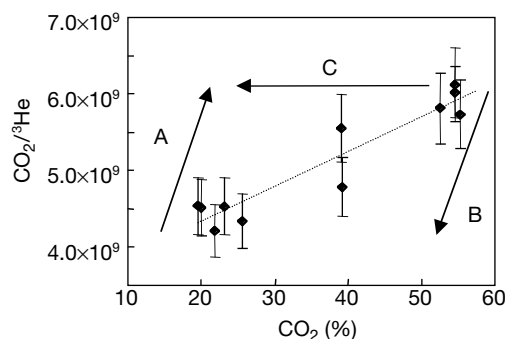


Figure 3 CO₂/He as a function of CO₂ content. CO₂/He in natural gases falls into three ranges¹: (1) CO₂/He > 1×10^{10} , where the CO₂ is dominated by crustal CO₂; (2) CO₂/He = $(1-10) \times 10^9$, which is the observed mantle range; and (3) CO₂/He < 1×10^9 , in which magmatic gases have lost CO₂ by precipitation or reduction. The JM-BB samples are within the magmatic range of $(1-10) \times 10^9$, and show a clear increase in CO₂/He with increasing CO₂ content. Vector A shows the effect of crustal CO₂ addition. Vector B shows the effect of CO₂ loss. Vector C shows the effect of CH₄ addition (simple mixing). Simple crustal CO₂ addition, precipitation or mixing with CH₄ cannot account for the data distribution.

Table 1 Sample location and composition

Sample identification			Gas isotope ratios						Gas concentrations (cc/cc)				
Well	Latitude (deg)	Longitude (deg)	$^3\text{He}/^4\text{He}$ (R/R _a)	$^{20}\text{Ne}/^{22}\text{Ne}$	$^{21}\text{Ne}/^{22}\text{Ne}$	$^{40}\text{Ar}/^{36}\text{Ar}$	$\delta^{13}\text{C}(\text{CO}_2)$ (‰)	$\delta^{13}\text{C}(\text{CH}_4)$ (‰)	^4He (10 ⁻⁴)	^{20}Ne (10 ⁻⁹)	^{40}Ar (10 ⁻⁵)	CO ₂	CH ₄
Turk No. 1A	-101.87472	30.40166	0.543(16)	9.75(8)	0.0515(5)	2330(150)	-2.77	-37.68	nm	nm	nm	0.548	0.446
Turk State No. 1	-101.85642	30.38758	0.527(16)	10.02(3)	0.0384(2)	2537(37)	-2.88	-37.48	1.25(9)	7.24(36)	2.87(43)	0.546	0.448
Bassett Goode No. 1	-101.83068	30.37852					-2.91	-37.53				0.551	0.443
Bassett Goode No. 3	-101.81341	30.37288	0.502(15)	9.72(3)	0.0432(3)	4811(160)	-2.89	-37.47	1.42(10)	5.12(26)	4.96(74)	0.553	0.442
Bassett Brown Unit No. 2	-101.8311	30.36766					-2.97	-37.62				0.574	0.421
Brown-Bassett No. 2*	-101.7995	30.34433	0.504(15)	9.61(5)	0.0614(6)	5660(160)	-2.9	-37.48	1.33(7)	2.14(11)	3.22(34)	0.546	0.449
Banner Hodge Unit No. 3	-101.78007	30.33781					-2.79	-37.46				0.542	0.452
Goode Est. C Unit No. 1	-101.76481	30.36819					-2.95	-37.32				0.397	0.595
Mayme K. Martin ETAL No. 1	-101.74721	30.35661	0.372(11)	nm	nm	1030(8)	-2.97	-37.3	1.42(10)	26.3(13)	4.32(65)	0.391	0.600
Mitchell 109 No. 1	-101.70934	30.33195					-2.91	-37.45				0.389	0.603
Mitchell 109 No. 2*	-101.69826	30.33329	0.400(12)	9.58(4)	0.073(6)	4575(85)	-2.92	-37.48	1.53(11)	1.67(8)	3.08(38)	0.392	0.601
Mitchell 4 No. 1X	-101.69417	30.32581					-2.88	-37.51				0.460	0.534
Mitchell 5 No. 1	-101.68429	30.32352	0.478(14)	9.95(4)	0.0371(2)	2322(46)	-2.84	-37.57	1.40(10)	9.58(48)	4.68(70)	0.525	0.468
Mitchell 103 No. 3	-101.64623	30.35543					-2.94	-37.38				0.208	0.768
Mitchell 103 No. 2	-101.63642	30.3568	0.246(7)	9.76(8)	0.0343(3)	2817(35)	-2.7	-37.31	1.39(10)	13.2(7)	3.09(46)	0.201	0.792
Mitchell No. 9	-101.60772	30.34996					-2.91	-37.32				0.210	0.783
Mitchell No. 21	-101.59929	30.35013					-2.97	-37.27				0.222	0.771
Mitchell No. 6	-101.58835	30.351	0.264(8)	9.28(14)	0.0733(11)	3554(55)	-2.96	-37.32	1.51(11)	1.65(8)	3.50(52)	0.219	0.774
Mitchell No. 3	-101.61307	30.33966	0.240(7)	9.37(3)	0.0623(3)	2970(49)	-3.06	-37.26	1.39(10)	2.38(12)	3.23(49)	0.197	0.798
Mitchell No. 2	-101.59188	30.34038					-2.97	-37.28				0.246	0.748
Mitchell No. 1	-101.5939	30.33159					-2.89	-37.31				0.261	0.733
Mitchell No. 8	-101.57924	30.32257					-2.99	-37.32				0.273	0.721
Mitchell 9 No. 13	-101.58623	30.3017					-2.97	-37.32				0.258	0.737
Mitchell A-11 No. 1	-101.57677	30.30286	0.272(8)	9.41(4)	0.0674(7)	3834(184)	-2.93	-37.3	1.66(12)	1.35(7)	3.26(49)	0.257	0.738
Mitchell No. 12	-101.57295	30.29118	0.267(8)	9.72(4)	0.0403(3)	2709(54)	-2.96	-37.32	1.46(10)	7.31(36)	3.54(53)	0.231	0.763
Bunger-Mitchum 46 No. 1	-101.56088	30.27417					-3.02	-37.27				0.211	0.783

Errors are 1 σ shown as last significant figures in parentheses. $\delta^{13}\text{C}(\text{CO}_2)$ error is 0.03‰ and $\delta^{13}\text{C}(\text{CH}_4)$ error is 0.02‰. nm, not measured.

* Average of two analyses for He and Ar.

seated magma body is consistent with both $\text{CO}_2/{}^3\text{He}$ and $\delta^{13}\text{C}(\text{CO}_2)$ variation (Fig. 4), and also with the magnitude of magmatic degassing inferred from resolved mantle ${}^3\text{He}/{}^{22}\text{Ne}$ in other natural gases⁵. This result suggests that the CO_2 in the JM-BB gas field is dominated by magmatic CO_2 with a negligible crustal contribution. Within the context of the magmatic degassing model, the JM-BB samples with the lowest $\text{CO}_2/{}^3\text{He}$ preserve the latest stages of outgassing and, assuming a simple filling history, will be the nearest to the magmatic source. In the JM-BB field the lowest $\text{CO}_2/{}^3\text{He}$ values are found in the samples with the lowest CO_2 content and are furthest from the MTB. This trend is reflected on the regional scale (Fig. 1), suggesting that the CO_2 source is to the north or east of the MTB.

Tertiary volcanism is associated with the basin and ranges province to the west of the Val Verde basin²³, and some 100 km away from the JM-BB study. Although a potential source of magmatic volatiles, this is consistent with neither the inferred direction of filling (Fig. 1) nor the magmatic ${}^3\text{He}$ associated with the CH_4 endmember (Fig. 2). We must then account for the increase in CO_2 content as the MTB is approached. We now consider CO_2 emplacement pre-dating CH_4 generation. CH_4 is generated in the hydrocarbon 'kitchens' in the basin deeps to the north of the gas fields. Assuming simple filling, the traps closest to these 'kitchens' would have the highest CH_4 content. Generation and migration of hydrocarbons within a system exposed to magmatic CO_2 flushing might be expected to result in a residual magmatic ${}^3\text{He}$ component in the hydrocarbon phase. This is what we observe. CO_2 charging, therefore, pre-dates the onset of hydrocarbon generation in the basin, which occurred about 280 Myr ago^{6,24}.

Magmatic volatile release is associated with Cenozoic extension^{11,25,26} and related basement uplift²⁵, as well as with deep faulting unassociated with clear extensional features²⁷. Fluids in the uplifted regions are often CO_2 -rich and contain magmatic ${}^3\text{He}$ as well as a strong crustal ${}^4\text{He}$ overprint²⁵. The JM-BB CO_2 endmember also has this character (Fig. 2). Maximum uplift of the Central basin and Ozona platforms occurred between 310 and 280 Myr ago in response to the MTB loading. Associated deep volatile release would provide the appropriate timing, mechanism and required spatial consistency to be the source of the magmatic CO_2 preserved in the Val Verde basin. Identification of magmatic CO_2 in the Val Verde foreland basin requires that we re-assess the role of magmatic CO_2

emplacement in basins not associated with recent rifting. The Rayleigh fractionation model proposed to account for regional $\text{CO}_2/{}^3\text{He}$ variation provides an important tool to identify the direction of magmatic CO_2 input into a basin system; this model also accounts for the higher $\text{CO}_2/{}^3\text{He}$ and heavier $\delta^{13}\text{C}(\text{CO}_2)$ often found in intracrustal manifestations of magmatic gas^{25,26} compared with the values in pristine mantle samples²⁸. Finally, the age of emplacement inferred from this study suggests that calculations of natural gas residence times based on diffusion experiments^{7,8} seriously underestimate the storage efficiency of some trapping structures. For the JM-BB field it can be shown that even individual compartments in the field were closed systems since about 300 Myr ago (M.S. *et al.*, manuscript in preparation). These observations provide support for the viability of natural gas exploration in deeper, older—and therefore more unconventional—locations. □

Received 5 April 2000; accepted 26 October 2000.

- Sherwood Lollar, B., Ballentine, C. J. & O'Nions, R. K. The fate of mantle-derived carbon in a continental sedimentary basin: Integration of C/He relationships and stable isotope signatures. *Geochim. Cosmochim. Acta* **61**, 2295–2307 (1997).
- Kharaka, Y. K., Carothers, W. W. & Rosenbauer, R. J. Thermal decarboxylation of acetic acid: implications for origin of natural gas. *Geochim. Cosmochim. Acta* **47**, 397–402 (1983).
- Hutcheon, I. & Abercrombie, H. Carbon dioxide in clastic rocks and silicate hydrolysis. *Geology* **18**, 541–544 (1990).
- Schoell, M. & Cathles, L. M. High CO_2 in natural gases as a late stage high temperature component in the evolution of petroleum systems. *Abstr. Am. Chem. Soc.* **215**, U623–U623 (1998).
- Ballentine, C. J. Resolving the mantle He/Ne and crustal ${}^{21}\text{Ne}/{}^{22}\text{Ne}$ in well gases. *Earth Planet. Sci. Lett.* **152**, 233–250 (1997).
- Horak, R. L. Tectonic and hydrocarbon maturation history in the Permian basin. *Oil Gas J.* **83**, 124–129 (1985).
- Schlomer, S. & Krooss, B. M. Experimental characterisation of the hydrocarbon sealing efficiency of cap rocks. *Mar. Petrol. Geol.* **14**, 563–578 (1997).
- Krooss, B. M., Leythaeuser, D. & Schaefer, R. G. The quantification of diffusive hydrocarbon losses through cap rocks of natural gas reservoirs - a reevaluation. *Am. Assoc. Petrol. Geol.* **76**, 403–406 (1992).
- Caffee, M. W. *et al.* Primordial noble gases from the Earth's mantle: Identification of a primitive volatile component. *Science* **285**, 2115–2118 (1999).
- Jenden, P. D., Hilton, D. R., Kaplan, I. R. & Craig, H. Abiogenic hydrocarbons and mantle helium in oil and gas fields. *Prof. Pap. US Geol. Surv.* **1570**, 31–56 (1993).
- Oxburgh, E. R., O'Nions, R. K. & Hill, R. I. Helium isotopes in sedimentary basins. *Nature* **324**, 632–635 (1986).
- Trull, T. W. & Kurz, M. D. Experimental measurements of ${}^3\text{He}$ and ${}^4\text{He}$ mobility in olivine and clinopyroxene at magmatic temperatures. *Geochim. Cosmochim. Acta* **57**, 1313–1324 (1993).
- Marty, B. & Jambon, A. $\text{C}/{}^3\text{He}$ in volatile fluxes from the solid Earth: implications for carbon geodynamics. *Earth Planet. Sci. Lett.* **83**, 16–26 (1987).
- Trull, T., Nadeau, S., Pineau, F., Polve, M. & Javoy, M. C-He systematics in hotspot xenoliths: Implications for mantle carbon contents and carbon recycling. *Earth Planet. Sci. Lett.* **118**, 43–64 (1993).
- Hilton, D. R., McMurtry, G. M. & Goff, F. Large variations in vent fluid $\text{CO}_2/{}^3\text{He}$ ratios signal rapid changes in magma chemistry at Loihi seamount, Hawaii. *Nature* **396**, 359–362 (1998).
- Burnard, P., Graham, D. & Turner, G. Vesicle-specific noble gas analysis of "popping rock": Implications for primordial noble gases in earth. *Science* **276**, 568–571 (1997).
- Schumaker, R. C. Paleozoic structure of the Central basin uplift and adjacent Delaware basin, West Texas. *Bull. Am. Assoc. Petrol. Geol.* **76**, 1804–1824 (1992).
- Ye, H., Royden, L., Burchfiel, C. & Schuepbach, M. Late Paleozoic deformation of interior North America: The Greater Ancestral Rocky Mountains. *Bull. Am. Assoc. Petrol. Geol.* **80**, 1397–1432 (1996).
- Montgomery, S. L. Val Verde Basin: Thrusted Strawn (Pennsylvanian) carbonate reservoirs, Pakenham field area. *Bull. Am. Assoc. Petrol. Geol.* **80**, 987–998 (1996).
- Ballentine, C. J., O'Nions, R. K., Oxburgh, E. R., Horvath, F. & Deak, J. Rare gas constraints on hydrocarbon accumulation, crustal degassing and groundwater flow in the Pannonian Basin. *Earth Planet. Sci. Lett.* **105**, 229–246 (1991).
- Kennedy, B. M., Hiyagon, H. & Reynolds, J. H. Crustal neon: a striking uniformity. *Earth Planet. Sci. Lett.* **98**, 277–286 (1990).
- Sackett, W. M. & Chung, M. H. Experimental confirmation of the lack of carbon isotope exchange between methane and carbon oxides at high temperatures. *Geochim. Cosmochim. Acta* **43** (1979).
- Henry, C. D., Price, J. G. & James, E. W. Mid-Cenozoic stress evolution and magmatism in the southern Cordillera, Texas and Mexico. Transition from continental arc to intraplate extension. *J. Geophys. Res.* **96**, 13545–13560 (1991).
- Ewing, T. E. in *New Mexico Geological Society Guidebook, 44th Field Conf.* (eds Love, D. W. *et al.*) 155–167 (New Mexico Geological Society, Socorro, 1993).
- Griesshaber, E., O'Nions, R. K. & Oxburgh, E. R. Helium and carbon isotope systematics in crustal fluids from the Eifel, the Rhine Graben and Black Forest, F. R. G. *Chem. Geol.* **99**, 213–235 (1992).
- Weinlich, F. H. *et al.* An active subcontinental mantle volatile system in the western Eger rift, Central Europe: Gas flux, isotopic (He, C, and N) and compositional fingerprints. *Geochim. Cosmochim. Acta* **63**, 3653–3671 (1999).
- Kennedy, B. M. *et al.* Mantle fluids in the San Andreas fault system, California. *Science* **278**, 1278–1281 (1997).
- Javoy, M. & Pineau, F. The volatile record of a 'popping' rock from the Mid-Atlantic Ridge at 14°N: chemical and isotopic composition of gas trapped in vesicles. *Earth Planet. Sci. Lett.* **107**, 598–611 (1991).
- Mattey, D. P. Carbon-dioxide solubility and carbon isotope fractionation in basaltic melt. *Geochim. Cosmochim. Acta* **55**, 3467–3473 (1991).
- Bottinga, Y. & Javoy, M. MORB degassing: bubble growth and ascent. *Chem. Geol.* **81**, 255–270 (1991).

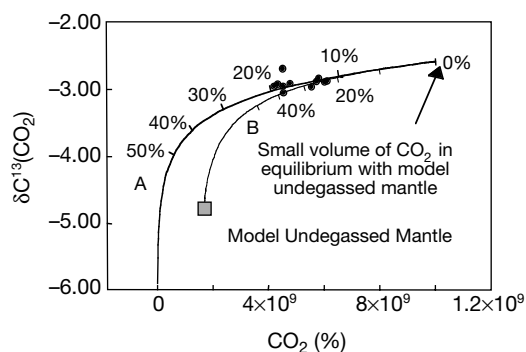


Figure 4 Evolution of $\delta^{13}\text{C}(\text{CO}_2)$ and $\text{CO}_2/{}^3\text{He}$ calculated for the gas phase of a degassing magma body. Two models are shown. Line A: the composition of the gas evolving from the magma by a Rayleigh fractionation process. Line B: the composition of this gas in an accumulating reservoir. The model undegassed magma is taken to have $\delta^{13}\text{C}(\text{CO}_2) = -4.7\text{‰}$ and $\text{CO}_2/{}^3\text{He} = 2 \times 10^9$, and is within the range estimated for the mantle source. The tick marks are the percentage loss of CO_2 from the magma body. $\delta^{13}\text{C}(\text{CO}_2)$ fractionation between a CO_2 gas phase and magma is taken to be 2‰ (ref. 29), and the relative solubility of $\text{He}/\text{CO}_2 = 5$ (ref. 30). Both $\text{CO}_2/{}^3\text{He}$ and $\delta^{13}\text{C}(\text{CO}_2)$ of the JM-BB field are consistent with this magma degassing model. Samples with the highest $\text{CO}_2/{}^3\text{He}$ are from the earliest stages of outgassing and correlate with CO_2 content (Figs 1 and 3).

Acknowledgements

We thank P. Jenden and M. Laroche for discussions, and L. Price and B. Marty for comments and suggestions. This work was supported by the US Department of Energy, the Gas Research Institute project 'Advanced stable isotope techniques' and the ETH, Zürich.

Correspondence and requests for materials should be addressed to C.J.B. (e-mail: ballentine@erdw.ethz.ch).

The electric Moho

Alan G. Jones^{*} & Ian J. Ferguson[†]

^{*} Geological Survey of Canada, Ottawa, Ontario, Canada, K1A 0E9

[†] Department of Geological Sciences, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2

Since Mohorovičić¹ discovered a dramatic increase in compressional seismic velocity at a depth of 54 km beneath the Kulpa Valley in Croatia, the 'Moho' has become arguably the most important seismological horizon in Earth owing to its role in defining the crust–mantle boundary. It is now known to be a ubiquitous feature of the Earth, being found beneath both the continents and the oceans, and is commonly assumed to separate lower-crustal mafic rocks from upper-mantle ultramafic rocks. Electromagnetic experiments conducted to date, however, have failed to detect a corresponding change in electrical conductivity

at the base of the crust, although one might be expected on the basis of laboratory measurements². Here we report electromagnetic data from the Slave craton, northern Canada, which show a step-change in conductivity at Moho depths. Such resolution is possible because the Slave craton is highly anomalous, exhibiting a total crustal conductance of less than 1 Siemens—more than an order of magnitude smaller than other Archaean cratons. We also found that the conductivity of the uppermost continental mantle directly beneath the Moho is two orders of magnitude more conducting than laboratory studies on olivine would suggest, inferring that there must be a connected conducting phase.

Earth materials conduct electricity predominantly by the flow of ions in fluids and the flow of electrons in solids. Dry silicate rocks are highly resistive, so a region of enhanced electrical conductivity represents an interconnected network of a fluid and/or mineral conducting phase. Laboratory studies suggest that for dry rock assemblages there may be an observable difference in conductivity between deep crustal mafic rocks and upper-mantle ultramafic rocks². However, a convincing step-change in conductivity at the base of the crust has not previously been reported because of one still inadequately explained feature of the Earth; the enhanced electrical conductivity of much of the continental lower crust³. This characteristic of the deep crust has been observed globally over the past 30 years, principally using the natural-source magnetotelluric (MT) method, and explanations for its existence remain controversial³. Suggestions of an interconnected brine below the brittle–ductile transition^{4,5} have been met with scepticism on petrological grounds⁶. A counter suggestion of an interconnected, thin, grain-boundary carbon film⁷ also has its critics⁸. Notwithstanding an explanation of its cause, a direct consequence of the existence of this lower-crustal conducting layer is that it is virtually impossible to determine its thickness, and hence derive total crustal thickness using electromagnetic (EM) methods.

Where the Earth consists of horizontal layers the external sources induce only horizontal electric currents in these layers, and the MT method is incapable of resolving independently the conductivity and thickness of a conductive layer sandwiched between two

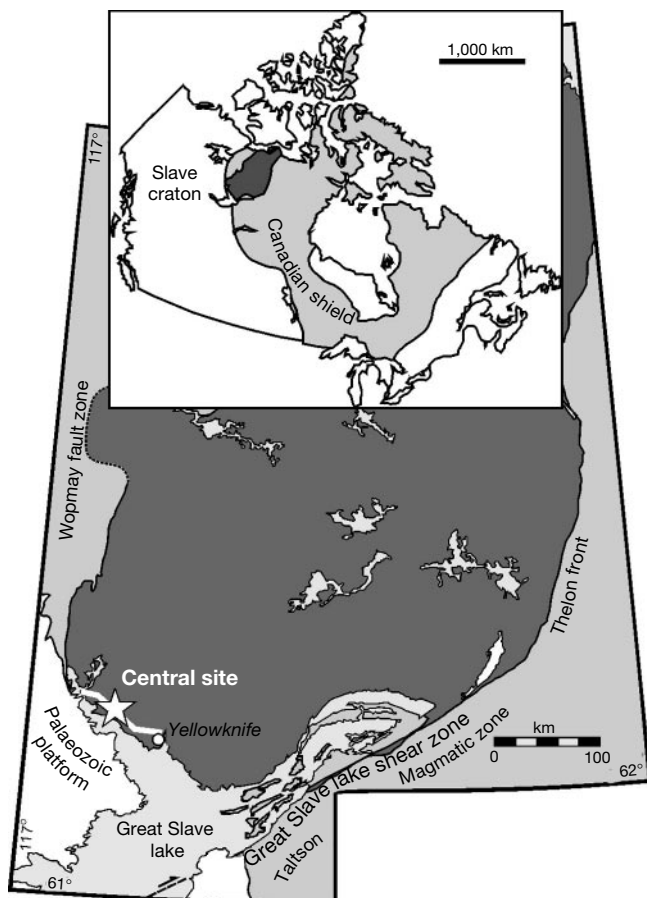


Figure 1 Tectonic map of the Slave craton in the northwestern part of the Canadian Shield. Also shown is the location of the profile and the central site (star).

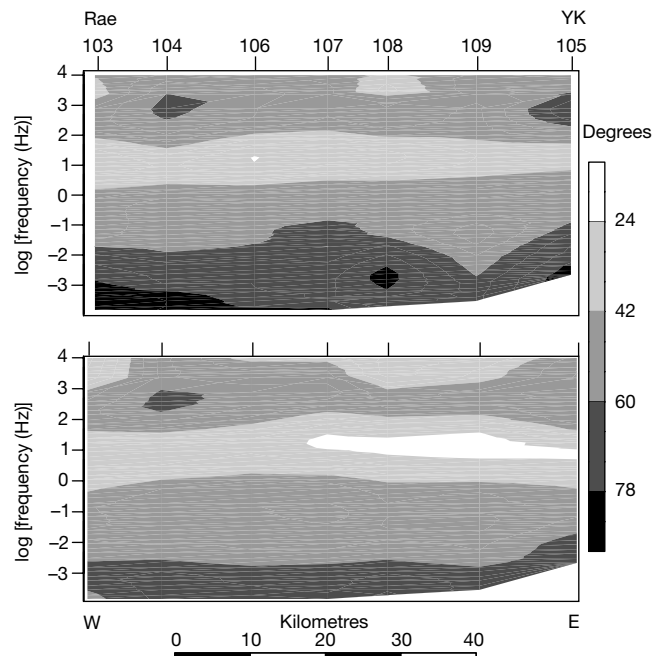


Figure 2 Contoured magnetotelluric phase responses for the seven sites from Yellowknife to the surface expression of the western boundary of the Slave craton. Distance is along the abscissa, and log(frequency) is along the ordinate. Top, phases with the electric component directed N41W. Bottom, phases with the electric component directed N49E.

resistive layers^{5,9}. Vertical electric currents induced in the Earth near laterally varying structures will be influenced (as a function of their wavelength) by both the conductivity and the thickness of the lower crust¹⁰. However, without independent knowledge of the exact conductivity structure of the upper-crustal features, it is again impossible to resolve the thickness of a conducting lower crust. Given these considerations, all interpretations of MT data that include a conducting lower crust and a step-change in conductivity at the base of the crust are suspect.

In almost all regions thus far studied, the presence of an enhanced electrically conducting layer in the lower crust screens the region directly below from detailed investigation by conventional MT experiments¹¹. In general, this effect limits the ability to resolve uppermost-mantle conductivity structure and only a maximum bound can be placed on its value¹¹. Only when the lower-crustal conducting layer is absent can the conductivity of continental uppermost mantle directly beneath the Moho be determined. Consequently, published values of uppermost-mantle conductivity¹², typically in the range 0.005–0.0125 S m⁻¹, must also be treated with caution. The conductivity of the deeper mantle below ~100 km has been defined by a number of other EM experiments; the greatest precision is that of ref. 13.

We conducted an MT survey as part of Lithoprobe's SNORCLE¹⁴ (Slave-northern Cordillera lithospheric evolution) transect; this survey involved acquisition of wide-band (10,000–0.01 Hz) and low-frequency (0.1–0.0001 Hz) data at locations on the south-western corner of the Archaean Slave craton (Fig. 1). The Slave craton, located in the northwestern Canadian Shield (Fig. 1 inset), is one of the world's smallest Archaean cratonic provinces (400 × 600 km), and hosts the oldest-known terrestrial rocks, the 4.03-Gyr-old Acasta gneisses¹⁵.

The MT sites on the exposed craton were along a 150-km-long east–west profile. The six sites west of Yellowknife (along line in Fig. 1) were on the Anton complex, an integral part of an early- to mid-Archaean basement complex covering much of the craton¹⁶. The time series data acquired at each site were processed using a robust multi-remote-reference algorithm (method 6 in ref. 17), and the estimated responses were corrected for local distortions of the

electric field¹⁸. In contrast to MT responses obtained on other exposed shields, distortions were unusually small at all sites, indicative of little near-surface conducting heterogeneity. The high-frequency responses for crustal depths (10,000–0.1 Hz) show weak evidence of a preferred geoelectric strike, whereas the lower-frequency responses, corresponding to signal penetration into mantle lithosphere, suggest a geoelectric strike of ~N41W geographical. The contoured MT phases in the frequency range 10,000–0.1 Hz for all six sites in the two orthogonal directions, with distance on the ordinate and log(frequency) on the abscissa, show uniform lateral behaviour (Fig. 2). This is characteristic of a region in which conductivity varies with depth alone within the crust and uppermost mantle, consistent with the lack of preferred strike orientation. The apparent resistivity curves (not shown) are close to one other, indicative of small amplitude distortion effects¹⁹.

The data from a central site on this profile, site 106, are representative of the whole profile and are shown in Fig. 3. Apart from scatter in the high-frequency 'dead-band' between 1 kHz and 3 kHz (ref. 20) and at the two lowest periods, the data are of excellent quality and the apparent resistivity and phase data are self-consistent²¹. Inverting the 27 averaged impedances²², expressed as apparent resistivities and phases, in the frequency range 1,000–0.1 Hz, with minimum assigned errors of 1° in phase and 3.5% in apparent resistivity, yields two models shown in Fig. 4. These two models represent end-member cases of possible acceptable models; the layered-Earth model²³ (solid line in Fig. 4) represents the model with the fewest number of uniform layers that fits the responses, whereas the continuous model (dashed line in Fig. 4) represents the smoothest model in terms of having the smallest conductivity gradient with depth²⁴. The models fit the observations to an r.m.s. misfit of 1.48 and 1.28 respectively, compared to a minimum possible misfit of 1.07 for a model of conductance spikes²⁵. The models are consistent in exhibiting a shallow (<1 km), low-conductivity, uppermost layer underlain by a more conductive layer to a depth of 2–3 km. Below this is a region of very low conductivity (<0.000025 S m⁻¹) to some tens of kilometres depth, beneath which is a moderately conductive (0.00025 S m⁻¹) homogeneous basal layer. Based on the layered-Earth model, the depth to the basal layer occurs at 35.8 ± 1.5 km. This interface depth is the second-best-resolved model parameter⁹, and the data most sensitive to it are the apparent resistivities at frequencies of 2–0.2 Hz and the phases at 10–1.25 Hz. Inverting only the 12 data at frequencies of

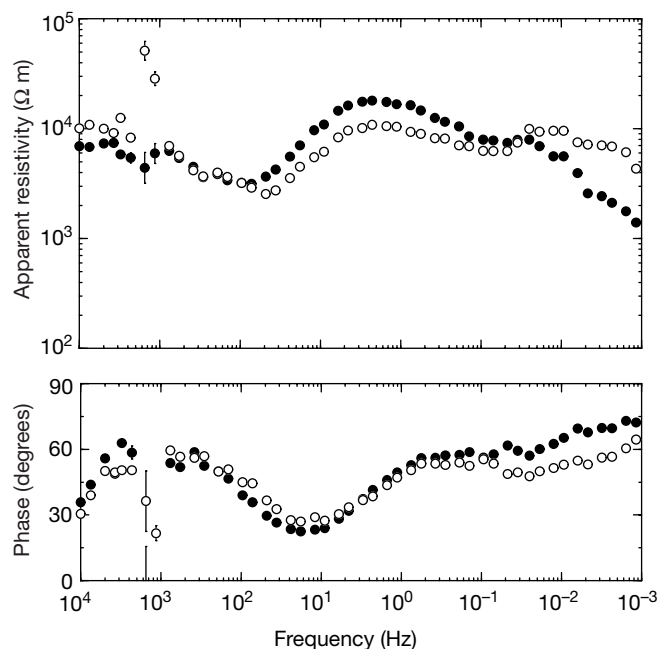


Figure 3 Magnetotelluric response from a central site on the profile, site 106 (Fig. 2). Filled circles, responses with the electric component directed N41W; open symbols, responses with the electric field directed N49E.

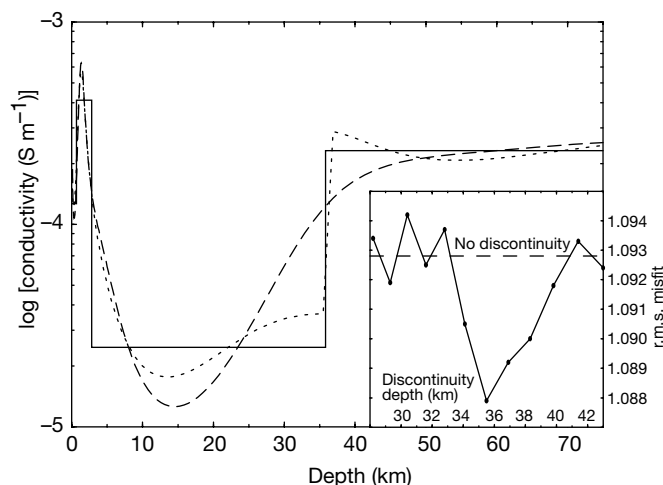


Figure 4 One-dimensional conductivity–depth models that fit the averaged impedances²¹ of the responses shown in Fig. 3. Solid line, best-fitting layered-Earth model. Dashed line, best-fitting continuous model. Dotted line, best-fitting continuous model with a permitted step change at 35.5 km. Inset, r.m.s. misfit for best-fitting continuous models varying the depth to the permitted step change.

10–0.5 Hz with the smooth model code²⁶ results in a much larger resistivity gradient at 30–40 km depth for the same r.m.s. misfit, implying that the data sensitive to that depth range are better fitted with a large gradient in conductivity.

Seismic reflection²⁶, refraction²⁷ and teleseismic²⁸ studies along the same profile are consistent in interpreting Moho depth at ~36 km. Thus, we can associate the conductivity change in the layered-Earth model with the seismically defined crust–mantle boundary. Re-performing the smooth model inversion but allowing a stepwise change in conductivity at 35.5 km yields the third model shown in Fig. 4. This model has a lower minimum misfit than the model without the step change and closely resembles the layered-Earth model below 10 km, featuring an order-of-magnitude increase in conductivity across the boundary. Varying the depth of the permitted step from 28 to 42 km (Fig. 4 inset) yields a minimum achievable r.m.s. misfit at a depth of 35.5 ± 1.4 km for both the full dataset (1,000–0.1 Hz), and the reduced range dataset (10–0.2 Hz).

This is, to our knowledge, the first definitive identification of a change in electrical conductivity at depths corresponding to the crust–mantle boundary, and the required resolution is a consequence of the absence of conducting material in the crust. It is also, to our knowledge the first time that the conductivity of the uppermost continental mantle directly beneath the Moho has been obtained, which also can be attributed to the lack of a conducting lower crust¹¹. The upper mantle beneath the Slave craton is two orders of magnitude more conducting than laboratory studies on olivine would suggest²⁹, inferring that there must be a connected conducting phase.

The conductivity for the lower crust of the Slave Province, at around $0.000025 \text{ S m}^{-1}$, is consistent with laboratory studies on candidate dry rock assemblages at appropriate facies conditions³⁰. In contrast, other mid- to late-Archaean cratons (Superior, Siberian, Baltic, Kaapvaal) have conducting material in their lower crust³, with a minimum conductance (conductivity-thickness product) of 20 S compared to less than 1 S for the Slave. Although the cause of the enhanced conductivity seen everywhere else remains contentious, what is clear is that whatever processes caused its existence elsewhere were not operating as crust was formed in the Slave craton. Recalling that the Slave craton hosts the oldest dated rocks, we speculate that this difference addresses questions of early Earth development and tectonic processes, and the applicability of plate-tectonic theory to the early- to mid-Archaean. If the enhanced conductivity in the continental lower crust of late-Archaean, Proterozoic and Palaeozoic terranes is emplaced by plate-tectonic processes of subduction and imbrication of sedimentary material (see, for example, ref. 31), then either such material was not available during the early Archaean or tectonic processes were operating differently. □

Received 22 June; accepted 8 November 2000.

- Mohorovičić, A. *Godišnje izvješće zagrebackog meteorološkog opservatorija za godinu 1909* 1–56 (Zagreb, 1910) (in Croatian); Das Beben vom 8. X. 1909. *Jahrb. Meteorol. Obs. Zagreb* 9, Teil 4, Absch. 1, 1–63 (1910) (in German); The earthquake from 8th October 1909. *Geofizika* 9, 3–55 (1992) (in English).
- Haak, V. in *Numerical Data and Functional Relationships in Science and Technology* Vol. 1, Subvol. b, Ch. 5, Sect. 4 (ed. Angenheister, G.) 291–307 (Springer, Berlin, 1982).
- Jones, A. G. in *Continental Lower Crust* Ch. 3 (eds Fountain, D. M., Arculus, R. & Kay, R. W.) 81–143 (Developments in Geotectonics 23, Elsevier, Amsterdam, 1992).
- Gough, D. I. Seismic reflectors, conductivity, water and stress in the continental crust. *Nature* 323, 143–144 (1986).
- Jones, A. G. MT and reflection: an essential combination. *Geophys. J. R. Astron. Soc.* 89, 7–18 (1987).
- Yardley, B. W. D. Is there water in the deep continental crust? *Nature* 323, 111 (1986).
- Frost, B. R. *et al.* Grain-boundary graphite in rocks and implications for high electrical conductivity in the lower crust. *Nature* 340, 134–136 (1989).
- Yardley, B. W. D. & Valley, J. W. The petrologic case for a dry lower crust. *J. Geophys. Res.* 102, 12173–12185 (1997).
- Jones, A. G. On the electrical crust-mantle structure in Fennoscandia: no Moho and the asthenosphere revealed? *Geophys. J. R. Astron. Soc.* 68, 371–388 (1982).
- Ferguson, I. J. & Edwards, R. N. Electromagnetic mode conversion by surface-conductivity anomalies; applications for conductivity soundings. *Geophys. J. Int.* 117, 48–68 (1994).
- Jones, A. G. Imaging the continental upper mantle using electromagnetic methods. *Lithos* 48, 57–80 (1999).

- Campbell, W. H. Introduction to electrical properties of the Earth's mantle. *Pure Appl. Geophys.* 125, 193–204 (1987).
- Schultz, A., Kurtz, R. D., Chave, A. D. & Jones, A. G. Conductivity discontinuities in the upper mantle beneath a stable craton. *Geophys. Res. Lett.* 20, 2941–2944 (1993).
- Clowes, R. M. (ed.) *LITHOPROBE Phase V Proposal* (Lithoprobe Secretariat, Univ. British Columbia, Vancouver, 1997).
- Stern, R. A. & Bleeker, W. Age of the world's oldest rocks refined using Canada's SHRIMP: The Acasta Gneiss Complex, Northwest Territories, Canada. *Geosci. Can.* 25, 27–31 (1998).
- Bleeker, W. *et al.* The Central Slave Basement Complex, Part I: its structural topology and autochthonous cover. *Can. J. Earth Sci.* 36, 1083–1109 (1999).
- Jones, A. G. *et al.* A comparison of techniques for magnetotelluric response function estimation. *J. Geophys. Res.* 94, 14201–14213 (1989).
- McNeice, G. & Jones, A. G. Multisite, multifrequency tensor decomposition of magnetotelluric data. *Geophysics* (in the press).
- Jones, A. G. Static shift of magnetotelluric data and its removal in a sedimentary basin environment. *Geophysics* 53, 967–978 (1988).
- Garcia, X. & Jones, A. G. Study of ionospheric sources for audiomagnetotelluric (AMT) sounding. *Geophysics* (submitted).
- Parker, R. L. & Booker, J. R. Optimal one-dimensional inversion and bounding of magnetotelluric apparent resistivity and phase measurements. *Phys. Earth Planet. Inter.* 98, 269–282 (1996).
- Berdichevsky, M. N. & Dmitriev, V. I. in *Geoelectric and Geothermal Studies* (ed. Adam, A.) 165–221 (KAPG Geophysical Monograph, Akademiai Kiado, Budapest, 1976).
- Fischer, G. & Le Quang, B. V. Topography and minimization of the standard deviation in one-dimensional magnetotelluric modelling. *Geophys. J. R. Astron. Soc.* 67, 279–292 (1981).
- Constable, S. C., Parker, R. L. & Constable, C. G. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics* 52, 289–300 (1987).
- Parker, R. L. The inverse problem of electromagnetic induction: existence and construction of solutions based on incomplete data. *J. Geophys. Res.* 85, 4421–4425 (1980).
- Cook, F. A. *et al.* Frozen subduction in Canada's Northwest Territories: Lithoprobe deep lithospheric reflection profiling of the western Canadian Shield. *Tectonics* 18, 1–24 (1999).
- Viejo, G. F., Clowes, R. M. & Amor, J. R. Imaging the lithospheric mantle in northwestern Canada with seismic wide-angle reflections. *Geophys. Res. Lett.* 26, 2809–2813 (1999).
- Bostock, M. G. Mantle stratigraphy and evolution of the Slave province. *J. Geophys. Res.* 103, 21183–21200 (1998).
- Constable, S., Shankland, T. J. & Duba, A. The electrical conductivity of an isotropic olivine mantle. *J. Geophys. Res.* 97, 3397–3404 (1992).
- Olhoeft, G. Electrical properties of granite with implications for the lower crust. *J. Geophys. Res.* 86, 931–936 (1981).
- Jones, A. G., Katsube, J. & Schwann, P. The longest conductivity anomaly in the world explained: sulphides in fold hinges causing very high electrical anisotropy. *J. Geomagn. Geoelectr.* 49, 1619–1629 (1997).

Acknowledgements

This work was performed under the auspices of Lithoprobe, Canada's national geoscience programme, funded by the Natural Sciences and Engineering Research Council of Canada and the Geological Survey of Canada. We thank the staff of Phoenix Geophysics Ltd for their attention to high-quality data acquisition.

Correspondence and requests for materials should be addressed to A.G.J. (e-mail: ajones@NRC.gc.ca).

Speciation in a ring

Darren E. Irwin*, Staffan Bensch* & Trevor D. Price

Department of Biology 0116, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA

The evolutionary divergence of a single species into two has never been directly observed in nature, primarily because speciation can take a long time to occur. A ring species, in which a chain of intergrading populations encircles a barrier and the terminal forms coexist without interbreeding, provides a situation in which variation in space can be used to infer variation in time^{1–3}. Here we reconstruct the pathway to speciation between two reproductively isolated forms of greenish warbler (*Phylloscopus trochiloides*). These two taxa do not interbreed in central Siberia but are connected by a long chain of intergrading populations encircling the Tibetan Plateau to the south⁴. Molecular data and climatic history imply that the reproductively

* Present address: Department of Ecology, Section of Animal Ecology, Lund University, Ecology Building, S-223 62 Lund, Sweden.

isolated taxa came into contact following expansions northward around the western and eastern sides of the plateau. Parallel selection pressures for increased song complexity during the northward expansions have been accompanied by divergence in song structure. Playback experiments show that the two Siberian forms do not recognize each other's songs. Our results show how gradual divergence in a trait involved in mate choice leads to the formation of new species.

The greenish warbler, *Phylloscopus trochiloides*, is a small (~7 g) insectivorous leaf-gleaning bird that breeds in forests over a range spanning much of the Palearctic (Fig. 1). In 1938 Ticehurst⁴ identified five subspecies encircling the treeless Tibetan Plateau that intergrade with each other, except in central Siberia to the north of the plateau (Fig. 1). On the basis of morphology and plumage patterns, both Ticehurst⁴ and Mayr¹ felt that the greenish warbler was a ring species, with evolution of reproductive isolation between the terminal forms resulting from spread around the Tibetan plateau.

Warblers in the genus *Phylloscopus* use song in species recognition^{5,6}. In a previous study⁷ we documented song variation in the greenish warbler by measuring spectrograms of songs

recorded at 15 sites throughout its geographic range. There is gradual variation in song characteristics around the ring (Fig. 2), but on both the west and east sides of the ring songs are longer and more complex in the north than in the south (Figs 1, 2). The songs of the putatively reproductively isolated forms (*viridanus* in the west and *plumbeitarsus* in the east) differ discontinuously in central Siberia. While *viridanus* constructs songs out of long song units with a high frequency range, *plumbeitarsus* constructs songs out of short units with low frequency range⁷ (Figs 1, 2).

We used song playback experiments to investigate the potential for reproductive isolation around the ring. Response by males to the playback of song is a widely used measure of whether different groups view each other as potential mates or competitors, and has been used to assign species status to several *Phylloscopus* taxa⁶. We conducted experiments in ten populations (PK, KL, PA, ML, MN and LN in the Himalayas, AA, YK, ST and AN further north) using recordings from other sites. Males responded strongly to recordings made up to 1,000–1,500 km away, but not to recordings made at further distances (Fig. 3). However, playbacks between *viridanus* and *plumbeitarsus* (shown by open circles in Fig. 3), even when separated by a relatively short distance, resulted in very little

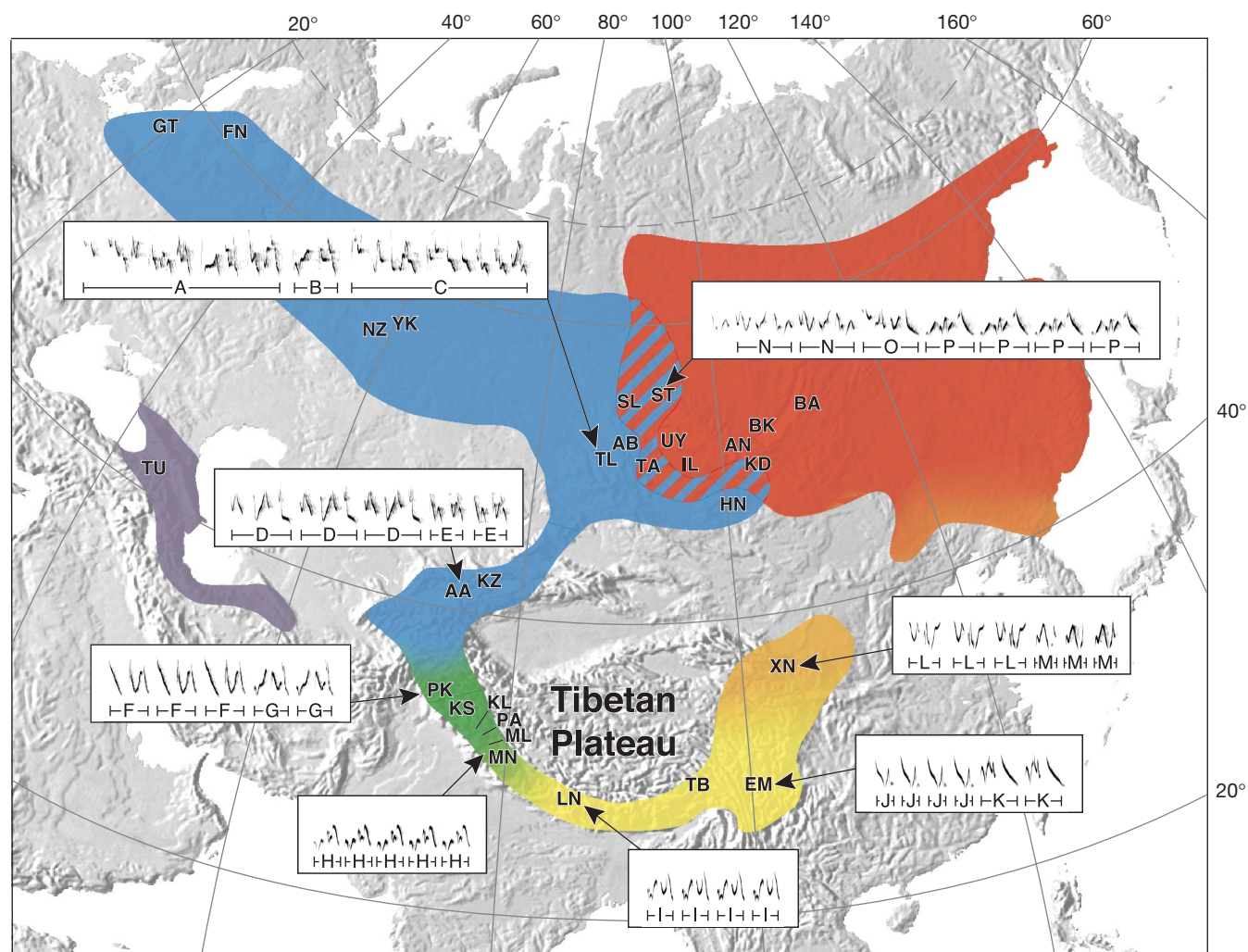


Figure 1 Geographic range of the greenish warbler species complex, along with research sites and representative song spectrograms. Different colours illustrate the ranges of six taxa commonly considered to be subspecies of *Phylloscopus trochiloides*⁴: purple, *nitidus*; blue, *viridanus*; green, *ludlowi*; yellow, *trochiloides*; orange, *obscuratus*; red, *plumbeitarsus*. Colours grade together in regions where Ticehurst⁴ described gradual change between subspecies. Research sites are indicated by their two-letter designation.

Also shown are representative song spectrograms (horizontal axis is time, vertical is frequency, darkness is amplitude) from eight locations⁷. Letters and brackets below the spectrograms indicate distinct song units. Song structure (for example, length of each unit, repetition of units, frequency range) differs between *viridanus* and *plumbeitarsus*, but there is a gradient in song around the southern side of the ring⁷.

response, indicating that the two taxa do not consider each other's song to be from their species. These results suggest that divergence of song and accompanying song recognition have been important in the evolution of reproductive isolation between the terminal forms of the ring.

We used molecular markers to reconstruct the biogeographical history of the ring. We constructed a mitochondrial DNA (mtDNA) gene tree for 149 individuals collected throughout the species' range, based on sequences of approximately 1,200 base pairs in the neighbourhood of the control region (Fig. 4). The gene tree

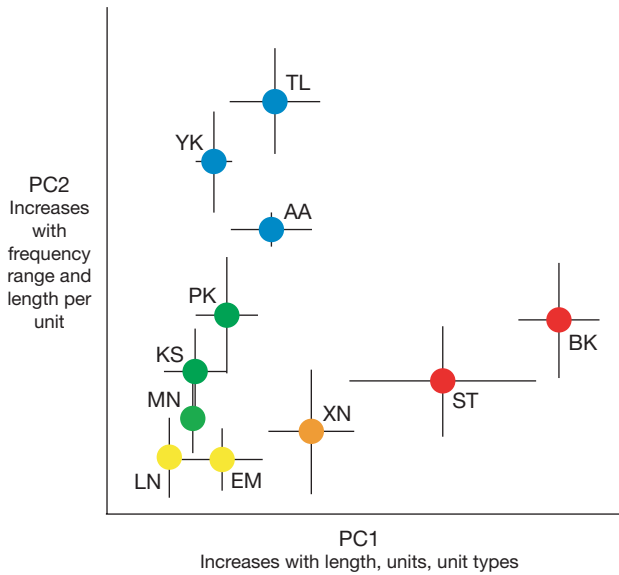


Figure 2 Geographic variation in the song of the greenish warbler as quantified by principal components analysis. Shown are population means and standard deviations, modified from ref. 7 with additional data. Songs distinctly differ between *viridanus* (blue) and *plumbeitarsus* (red), but change gradually through populations to the south (in order around the ring: YK-TL-AA-PK-KS-MN-LN-EM-XN-BK-ST). Both PC1 and PC2 are axes of complexity, and Himalayan populations (yellow and green) have the simplest songs (low PC1 and PC2).

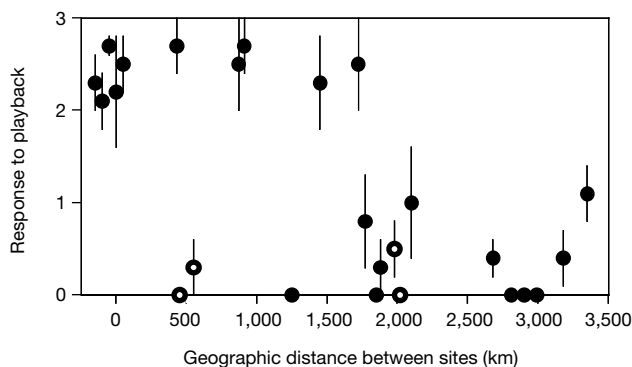


Figure 3 Relationship between geographic distance and song recognition. The horizontal axis shows the distance between the population in which song recordings were made and the population in which those recordings were used in a playback experiment. The vertical axis shows the response (mean $\pm$ s.e.) to those recordings. Open circles are playbacks of *viridanus* recordings to *plumbeitarsus* populations or vice versa, and filled circles are all other experiments. The cluster of points at a geographic distance of zero represent trials in which songs recorded in a population were played to birds in that population (although never to the same bird as the recording or his neighbour). Generally, birds respond strongly to recordings from populations up to 1,500 km away, but *viridanus* and *plumbeitarsus* across Siberia do not respond to each other even when only 500 km apart. Number of males tested for each point varied from 2 to 15 (mean 5.25).

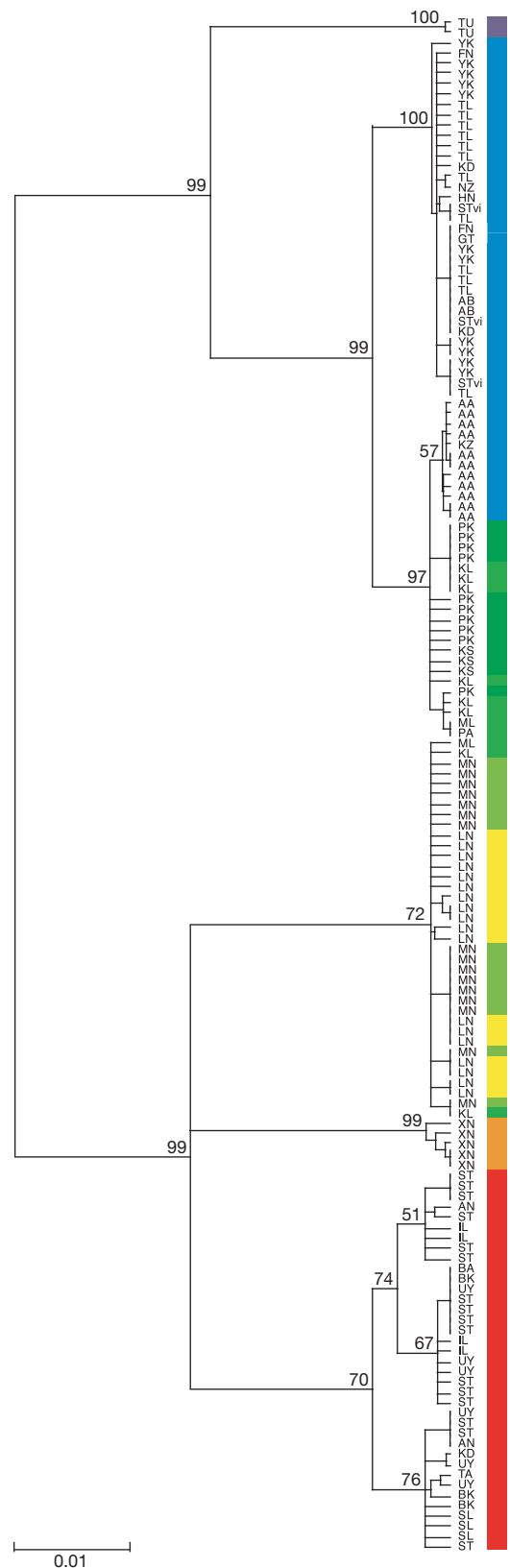


Figure 4 Mitochondrial DNA gene tree based on variation in $\sim 1,200$ bp in the neighbourhood of the control region and ND6 gene. The two-letter designation and colour bar at each tip of the tree indicate the research site the individual is from and the taxon to which it belongs (the three STvi individuals were *viridanus* males at site ST; the other ST individuals were all *plumbeitarsus*). We were unable to obtain sequence data for sites TB and EM. Numbers at the nodes are support values in per cent for major clades. The tree was produced using maximum-likelihood quartet puzzling²¹ assuming clock-like branch lengths (see Methods). Scale bar represents the substitution rate along a lineage.

shows strong geographic structure. Two major clades correspond to western and eastern individuals. We sampled intensively across about 500 km in each of the regions where the two clades meet. In central Siberia the concordance between haplotype (western or eastern clade) and song (*viridanus* or *plumbeitarsus*) was perfect (17 western birds and 35 eastern birds), providing no evidence for mitochondrial introgression between the two Siberian taxa. In the south, the region where the two mitochondrial clades meet does not correspond with a subspecies boundary, but instead occurs within the range of *ludlowi* (Figs 1 and 4). Songs of individuals carrying the different haplotypes are not distinguishable (for example, compare the similarity of KS and MN in Fig. 2), and playback experiments indicate that birds do not distinguish among them in this region. For example, birds at PK responded to MN song, even though the two are from different mitochondrial clades. And in the two populations where birds of both mitochondrial clades have been found to co-occur (KL, ML), individuals respond strongly to songs recorded at both PK (western haplotype) and MN (eastern haplotype) ($n = 10$ playback experiments).

Genetic variation in two microsatellite markers (Pocc2 and Pocc6) matches that of the mtDNA tree (Fig. 5a, b). For both microsatellites, mean length varies significantly over the range (analysis of variance (ANOVA) across localities: Pocc2, $F_{8,181} = 4.4$, $P < 0.0001$; Pocc6, $F_{8,191} = 21.6$, $P < 0.0001$). The central Siberian forms at TL and ST differ in mean allele length (*a posteriori* tests: Pocc2, $P < 0.01$; Pocc6, $P < 0.0001$). There is also significant variation in the south between Kashmir and Nepal (comparison of KS to LN: Pocc2, $P < 0.01$; Pocc6, $P < 0.0001$) but the central population in northwest India (MN) is clearly intermediate to the Kashmir and Nepal populations, suggesting continuing gene flow in that region. There is little variation from south to north along either side of the ring.

The microsatellite data and the mtDNA tree are consistent in showing high genetic change both in the north, between the central Siberian taxa, and in the south, across the western Himalayas. Song variation, the mtDNA tree and the microsatellite data together indicate that there is current gene flow across the southern part of the range, but not the northern. These results suggest two possible biogeographical histories. In the first, an ancestral species was split into two populations: one to the west and the other to the east.

These populations then expanded and met each other in the south, where they interbred, and in the north, where they did not interbreed. In the second model, an ancestral species was confined to the Himalayas and developed some genetic structure as a result of isolation by distance. The species then expanded northwards along two pathways: one from the western side of its range into central Asia and west Siberia, and one from the eastern side into China and east Siberia. Simulations (D.I., unpublished data) indicate that both biogeographical histories are consistent with the data. In particular, strong geographic structuring of mitochondrial gene trees readily arises in isolation-by-distance models, and cannot necessarily be taken as indication of past geographic separation of populations. With either model, it is clear that the Siberian populations have resulted from recent northward expansions, because ice ages have rendered most of Siberia treeless for long periods through the Pleistocene⁸.

The greenish warbler, like other ring species^{3,9,10}, has probably had a complex biogeographical history involving allopatric populations and secondary contact. Indeed, currently there is a gap in distribution in northeast China that is probably due to almost complete destruction of forests in that region by humans over the past several millennia^{11,12}. The recentness of this break is indicated by the relatively little differentiation in genetics, song and plumage across the gap (Figs 2, 4 and 5), and birds in east Siberia ($n = 3$) responded strongly to songs recorded in central China. The presence of past geographical breaks does not detract from the basic value of ring species as indicators of the way in which continuous variation within species is transformed to discontinuities between them^{3,9}. Around the greenish warbler's ring the two reproductively isolated forms in Siberia are connected by continuous variation in all measured phenotypic traits, including songs (Fig. 2), plumage patterns (Fig. 5c) and body size (Fig. 5d).

Two of the measured phenotypic traits (songs and plumage patterns) are used in courtship and territory defence^{7,13,14}. Both of these traits are distinctly different between the two reproductively isolated forms in Siberia. In the case of song, a thorough analysis of the geographic variation implicates reduced natural selection costs and/or increased sexual selection pressures as the cause of parallel evolution of greater complexity towards the north on both sides of the Tibetan Plateau⁷. The increased complexity has

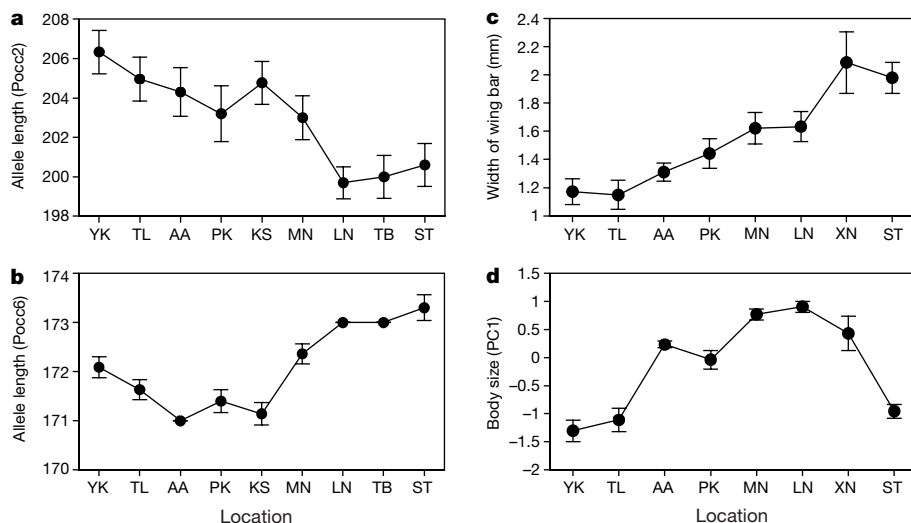


Figure 5 Variation in four traits (mean $\pm$ s.e.) around the ring of greenish warbler populations. Populations (see Fig. 1) are arranged in geographical order from west Siberia through the Himalayas to east Siberia. **a, b**, The lengths of two microsatellite loci (9–14 individuals per site); **c**, width of the greater covert wing bar (4 males at XN, 6 at PK, 10 at

all others); **d**, body size (4–28 males per site). Populations TL (*viridanus*) and ST (*plumbeitarsus*), both in central Siberia, are among the most divergent of all populations in the two microsatellites and wing bar. They are, however, similar in body size.

been accomplished in different ways on the west and east sides of the ring, resulting in song divergence⁷. A plumage characteristic, the size of the pale wing bar, also differs between populations, being smaller in west Siberia than in east Siberia (Fig. 5c; ANOVA across localities: $F_{7,62} = 9.6$, $P < 0.0001$; *a posteriori* comparison of TL and ST: $P < 0.0001$). Differences in wing-bar size among the *Phylloscopus* species have been related to habitat¹³. However, the habitat occupied by the reproductively isolated taxa in central Siberia is very similar, and an explanation for the divergence in wing bars remains to be uncovered.

In contrast to the patterns shown by song and plumage, body size is similar between the two reproductively isolated taxa. Differences in body size among species in the genus *Phylloscopus* have been related to ecological traits such as food type and feeding methods^{14,15}. Body size varies significantly around the ring ($F_{7,98} = 43.5$, $P < 0.0001$; Fig. 5d) but differs little between the northerly populations (*a posteriori* tests, comparison of TL and ST: $P = 0.44$). Apparently, parallel natural selection pressures across the west and east sides of the species range have resulted in parallel evolution of body size, from the larger southern type to the smaller northern types. Ecological similarity of the two reproductively isolated forms in the north is also suggested by the relatively small penetration of each into the other's range.

Much recent interest has been directed towards a possible role of natural selection in speciation^{16,17}, but sexually selected traits are most obviously divergent between the two reproductively isolated taxa in Siberia. It appears that by magnifying small differences in southern populations into species-level differences between the Siberian taxa, sexual selection has had a prime role in speciation. Sexual selection has been invoked as a powerful force in speciation partly because sexual selection often favours novelty and/or complexity, resulting in many possible outcomes¹⁸. We suggest that similar selection pressures for greater song complexity in both northward expansions has resulted in song divergence largely because there are many different ways to evolve greater complexity. As illustrated here for body size, natural selection may often lead to parallel or convergent evolution, resulting in little difference between incipient species when they meet. □

Methods

Fieldwork

We visited most sites for several weeks during at least one breeding season (May through early July). Sites TU, FN, NZ, KZ, TB, EM, HN, KD and BA were visited by others. Some sites were visited more than once (KS, six seasons; MN, four seasons; LN, two seasons; and ST, four seasons).

Collection of samples for molecular analysis

To obtain most of the DNA samples, we captured each bird in a mist net, collected a few drops of blood, and released the bird. We also obtained 12 skin samples from the British Museum (collected at site TB in the years 1922 to 1947) and 7 muscle samples from the Burke Museum (one each from sites NZ, KZ, BA and HN, and three from KD). DNA was extracted from samples using the QIAamp Tissue Kit (Qiagen).

Mitochondrial DNA gene tree

Because of a gene rearrangement¹⁹, *Phylloscopus* warblers have an unusual mitochondrial gene order surrounding the control region (cyt b, tRNA^{Thr}, control region, tRNA^{Pro}, ND6, tRNA^{Glu}, non-coding region, tRNA^{Phe}, 12S rRNA). We amplified a fragment from this region using the primers DLL3 (control region) and 12SH2 (12S rRNA). This fragment was sequenced with an ABI 377 DNA Sequencer at the UCSD Cancer Center in two reactions, using primers DLL3 and DLLF2 (5'-GGATCAGCTGCTAACGACAC-3'), which is located near the centre of the fragment. Sequences ranged in size from 1,006 to 1,298 bp, both because of insertions/deletions and because some sequences were incomplete at the ends. We aligned the sequences using the program Clustal W²⁰. To construct the gene tree, we used the program PUZZLE 4.0.2 (ref. 21) using an HKY model of substitution, parameters estimated from the data set (by the approximate method) and gamma-distributed rates (gamma-distribution parameter $\alpha = 0.32$). Gaps (from insertions/deletions) were treated as missing data. Out of the 149 individuals sequenced, there were 101 unique haplotypes and only these were used to construct the tree. The tree assuming a molecular clock (Fig. 4) is not significantly less likely than the no-clock tree (likelihood ratio test, $\chi^2_{99} = 99.35$, $P = 0.47$). Aligned sequences from all samples have been

deposited in GenBank (accession numbers AF316171–AF316319) and can also be obtained from the authors.

Microsatellites

We genotyped individuals for two microsatellite loci, Pocc2 and Pocc6 (ref. 22). We labelled primers with ³²P or ³³P, and used PCR conditions as in ref. 22. Products were resolved on Sequagel XR (National Diagnostics) polyacrylamide gels and exposed to film for 1–72 h.

Wing bar size

We collected the fourth greater covert feather from the right wing of most of the birds that we caught. Using a microscope, we measured the length of pale (non-melanized) colour along the shaft of the feather at the feather's tip (males only).

Body size

For every bird we caught, we measured six morphological traits¹⁵ (tarsus length, wing length, tail length, and beak length, depth and width). These variables were used in a principal components analysis (males only). The first principal component (PC1) weights all variables about equally, and is used as a measurement of body size.

Playback experiments

We prepared each playback tape by recording a singing male for 10 min. Whenever possible, three recordings (each from a different bird) from a source population were used in the playback experiments. At each target population, we conducted playbacks by locating a singing male, placing a speaker directly below the tree he was in, randomly determining which source tape to play (choosing from three tapes of each possible source population), and playing the tape for 7–10 min. We judged the response of the target bird on a scale of 0 (no response) to 3 (strong response, aggressively approaching the speaker).

Received 10 October; accepted 2 November 2000.

- Mayr, E. *Systematics and the Origin of Species* (Dover, New York, 1942).
- Wake, D. B., Yanev, K. P. & Frelow, M. M. in *Speciation and its Consequences* (eds Otte, D. & Endler, J.) 134–157 (Sinauer, Sunderland, MA, 1989).
- Wake, D. B. & Schneider, C. J. Taxonomy of the plethodontid salamander genus *Ensatina*. *Herpetologica* **54**, 279–298 (1998).
- Ticehurst, C. B. A *Systematic Review of the Genus Phylloscopus* (Johnson Reprint, New York, 1938).
- Martens, J. M. in *Ecology and Evolution of Acoustic Communication in Birds* (eds Kroodsma, D. E. & Miller, E. H.) 221–240 (Cornell Univ. Press, Ithaca, 1996).
- Irwin, D. E., Alström, P., Olsson, U. & Benowitz-Fredericks, Z. M. Cryptic species in the genus *Phylloscopus* (Old World leaf warblers). *Ibis* (in the press).
- Irwin, D. E. Song variation in an avian ring species. *Evolution* **54**, 998–1010 (2000).
- Frenzel, B. The Pleistocene vegetation of northern Eurasia. *Science* **161**, 637–649 (1968).
- Mayr, E. *Populations, Species, and Evolution* (Harvard Univ. Press, Cambridge, 1970).
- Highton, R. Is *Ensatina eschscholtzii* a ring-species? *Herpetologica* **54**, 254–278 (1998).
- Wang, C.-W. *The Forests of China* (Maria Moors Cabot Foundation Publication No. 5, Botanical Museum, Harvard Univ., Cambridge, MA, 1961).
- Menzies, N. K. *Forest and Land Management in Imperial China* (St. Martin's, New York, 1994).
- Marchetti, K. Dark habitats and bright birds illustrate the role of the environment in species divergence. *Nature* **362**, 149–152 (1993).
- Marchetti, K. & Price, T. The adaptive significance of colour patterns in the Old World leaf warblers, genus *Phylloscopus*. *Oikos* **79**, 410–412 (1997).
- Price, T. Morphology and ecology of breeding warblers along an altitudinal gradient in Kashmir, India. *J. Anim. Ecol.* **60**, 643–664 (1991).
- Orr, M. R. & Smith, T. B. Ecology and speciation. *Trends Ecol. Evol.* **13**, 502–506 (1998).
- Rundle, H. D., Nagel, L., Boughman, J. W. & Schluter, D. Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**, 306–308 (2000).
- West-Eberhard, M. J. Sexual selection, social competition and speciation. *Quart. Rev. Biol.* **58**, 155–183 (1983).
- Bensch, S. & Härlid, A. Mitochondrial gene rearrangements in songbirds. *Mol. Biol. Evol.* **17**, 107–113 (2000).
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
- Strimmer, K. & von Haeseler, A. Quartet-puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969 (1996).
- Bensch, S., Price, T. & Kohn, J. Isolation and characterization of microsatellite loci in a *Phylloscopus* warbler. *Mol. Ecol.* **6**, 91–92 (1997).

Acknowledgements

We thank J. Kohn for the use of his laboratory; P. Alström, K. Marchetti, U. Olsson, A. Richman, J. Tiainen, the British Museum and the Burke Museum for samples; Z. Benowitz-Fredericks, J. Gibson, S. Gross, J. Irwin, G. Kelberg, A. Knorre, K. Marchetti and B. Sheldon for help in the field; A. Asbeck, M. Bouvier, H. Neville, K. Petren, R. Radtkey and A. Richman for technical assistance; T. Case, J. Coyne, M. Dantzker, D. Holway, J. Irwin, J. Kohn, T. Pärt, A. Qvarnström, A. Suarez, N. Tsutsui, M. Turelli, A. Uy and S. Vehrencamp for comments on the manuscript. For financial support we thank the American Ornithologists' Union, The Explorers Club, the Jeanne Messier Memorial Fund, the National Geographic Society, the National Science Foundation and Sigma Xi.

Correspondence and requests for materials should be addressed to D.E.I. (e-mail: dirwin@biomail.ucsd.edu).

Three-butterfly system provides a field test of müllerian mimicry

Durrell D. Kapan*

Centre for Biodiversity Research, Department of Zoology, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada, and Section of Integrative Biology, Patterson Laboratories, University of Texas, Austin, Texas 78712-1064, USA

* Present address: Departamento de Biología, Universidad de Puerto Rico—Rio Piedras, San Juan 00931-3360, Puerto Rico

In 1879, Müller proposed that two brightly coloured distasteful butterfly species (co-models) that share a single warning-colour pattern would benefit by spreading the selective burden of educating predators^{1–5}. The mutual benefit of sharing warning signals among distasteful species, so-called müllerian mimicry, is supported by comparative evidence^{2,3}, theoretical studies^{5,6} and laboratory simulations⁷; however, to date, this key exemplar of adaptive evolution has not been experimentally tested in the field. To measure natural selection generated by müllerian mimicry, I exploited the unusual polymorphism of *Heliconius cydno* (Lepidoptera: Nymphalidae)⁸. Here I show increased survival of *H. cydno* morphs that match locally abundant monomorphic co-model species. This study demonstrates müllerian mimicry in the field. It also shows that müllerian mimicry with several co-models generates geographically divergent selection, which explains the existence of polymorphism in distasteful species with warning coloration⁹.

The importance of warning coloration in *Heliconius* butterflies has been examined by marking out wing colour-pattern elements or by transferring colour-pattern races across an inter-racial hybrid zone^{10,11}. These studies showed that there is strong, positive frequency-dependent (aposematic) selection¹² acting on colour pattern in a species, but failed to measure directly the benefit of müllerian mimicry. This is because they measured purifying selection on colour pattern in a numerically dominant co-model species *Heliconius erato* (W. W. Benson, personal communication)^{10,11,13} rather than selection for colour-pattern convergence between species (Fig. 1). The results of both studies demonstrate how natural selection operates on intraspecific variation in warning coloration (uninitiated predators rapidly eliminate rare warning-

colour variants within a species⁹), and can explain the observation that distasteful *Heliconius*^{14,15} species are generally monomorphic within a given geographic area⁹.

This study tests the evolution of müllerian mimetic coloration (in *Heliconius*) by exploiting an unusual system of colour-pattern polymorphism in one species of *Heliconius* butterfly. In western Ecuador, two colour morphs of the unpalatable *H. cydno* butterfly (yellow and white) resemble two different unpalatable species, *Heliconius eleuchia* (yellow) and *Heliconius sapho* (white) (Fig. 1; the proposed co-models of *H. cydno*⁸). Where all three species occur together, the two monomorphic co-models vary in relative abundance (see Methods)⁸.

I used the distasteful polymorphic *H. cydno* to test two predictions of müllerian mimicry theory: first, that unfamiliar experimental *H. cydno* morphs whose wing colour does not match the locally dominant co-model suffer more attacks by predators compared with control morphs whose wing colour does match the co-model¹; second, that differences in survival between experimental and control butterflies decreases with increasing release density. Predators must attack a number of butterflies to learn to avoid an unfamiliar experimental colour pattern¹⁵. These attacked butterflies will represent a lower fraction of the experimental butterflies when more experimental butterflies are released^{3,10,11}. In fact, predator education may have thwarted previous warning-colour experiments that failed to detect differences between experimental and control treatments because, in an attempt to achieve significant samples sizes, they used high-density releases, long release times or re-released butterflies at the same study site^{10,11,16–18}.

To test the first prediction, I captured pairs of yellow and white *H. cydno* butterflies from two source sites and released them at four other sites that were dominated by yellow or white co-models (Table 1, *H. eleuchia* or *H. sapho*). In contrast to previous experiments, this reciprocal transplant design measures the benefit that a rare unpalatable morph of *H. cydno* derives from müllerian mimicry with either of its more common co-model species (Fig. 1). I tested simultaneously the second prediction by releasing these *H. cydno* pairs at different densities: low density (1 pair every 163–194 m of trail) in three sites (Table 1, Manta Real, Agua Caliente and Tinalandia) and high density (1 pair every 40 m of trail) at one site (Table 1, Maquipucuna; also see Methods)⁸.

As predicted, experimental individuals had a lower frequency of being refound on subsequent visits (resighted) compared with controls for each study site (Fig. 2). From resight data, I used

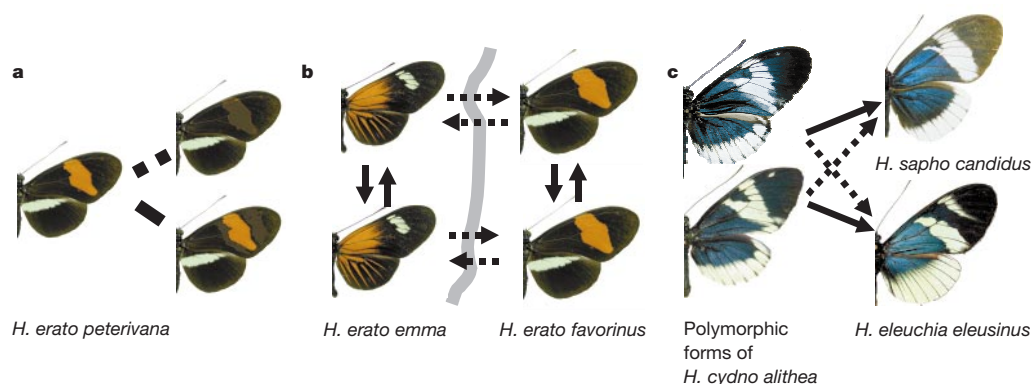


Figure 1 Experimental designs of aposematic selection and müllerian mimicry in *Heliconius* butterflies. **a**, Experimental (dashed line) and control butterflies (solid lines) of a single Costa Rican population of *H. erato peterivana*. The red forewing patch is darkened for experimentals and an equivalent amount of ink is applied to dark parts of the forewing for controls¹⁰. **b**, Experimental (dashed arrows) and control (solid arrows) butterflies of *H. erato emma* and *H. erato favorinus*. Experimental butterflies were transferred across their common hybrid zone (grey line) and control butterflies were transferred parallel to the

hybrid zone¹¹. In both studies abundant *H. erato* were manipulated at sites where they occurred with the rare co-model *H. melpomene* (W. W. Benson, personal communication)^{10,11,30}. **c**, Reciprocal transfer of polymorphic *H. cydno*. White *H. cydno* are control butterflies (solid arrow) when transferred to sites dominated by the white co-model (*H. sapho*) and experimental butterflies (dashed arrow) when transferred to sites dominated by the yellow co-model (*H. eleuchia*). The opposite is true for the yellow *H. cydno* butterfly.

maximum likelihood methods^{19,20} to estimate survival differences between experimental and control butterflies²¹. This involved partitioning butterfly disappearance into two phases: disappearance during the first 24 h (to estimate the probability of establishment, P_E) and subsequent disappearance (to estimate the exponential mortality rate, λ) (see Methods)¹¹. Disappearance rates of released butterflies were in the predicted direction: experimental butterflies had higher initial disappearance rates (lower P_E) and higher subsequent disappearance rates (Fig. 3, λ).

The greater disappearance rate of experimental butterflies is not due to emigration from release sites, as the dispersal distances of experimental and control butterflies were very similar. The mean maximum distance moved by each resighted butterfly did not differ between experimental (105 ± 35.9 m; mean $\pm$ s.e.m.) and control (118 ± 37.4 m) butterflies ($t_{41} = 0.41$, $P = 0.68$; on ln-transformed distance data). Very few released butterflies flew long distances. Only visually oriented predators are expected to affect experimental butterflies while leaving the controls alone. It is therefore reasonable to attribute differences in disappearance rate to differences in survival between morphs.

Morphs differed in survival (life expectancy, P_E/λ): the maximum likelihood estimate of life expectancy across all sites was 5 days for experimental butterflies and 14 days for control butterflies (Fig. 3b, $G_4 = 10.72$, $P = 0.03$). These differences in life expectancy for *H. cydno* butterflies, which have a fairly constant reproductive

output²², are equivalent to a selective coefficient (s) of 0.64 against experimentals (see Methods)¹¹.

Release density affects survival differences. At high release density, selection was not detected: life expectancies for experimental and control butterflies were similar (16 and 17 days respectively, $G_4 = 1.72$, $P = 0.79$). In contrast, life expectancy was 2 days for experimental and 12 days for control butterflies at the three sites of low-density release combined (Fig. 3a, $G_2 = 8.92$, $P = 0.012$).

Differences in life expectancy between experimental and control butterflies vary predictably with the density of release, suggesting that avian predators such as Jacamars (Galbulidae), motmots (Motmotidae) and tyrant flycatchers (Tyrannidae) are capable of learning and can select against unfamiliar coloration of experimental butterflies in the field¹⁵. The lower life expectancy of experimental relative to control butterflies at sites of low-density release indicates that looking like an abundant co-model increases the probability of survival. In contrast, at high release density, life-expectancy differences between treatments are negligible. Predators attack a number of experimental colour morphs to learn avoidance of an unfamiliar colour pattern, but at high release density this number will be a lower fraction of the total experimentals released. Even if predators are responsible for the high initial loss of experimental butterflies (low P_E) across all sites (Fig. 3), they would be expected to learn to avoid experimental morphs when frequent, accounting for the decreased difference between experimental and control λ observed at the high-density release site, Maquipucuna.

There are several implications of this study. The benefit of müllerian mimicry is very high. Predators quickly eliminate rare phenotypes that deviate from the local fitness peak provided by abundant co-models. Given this, polymorphism in distasteful

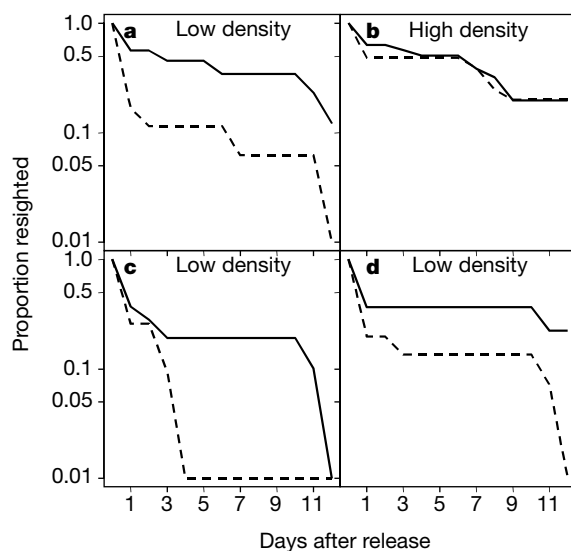


Figure 2 Observed proportion (log₁₀ scale) of control and experimental *H. cydno* butterflies resighted after initial release. Control (solid line) and experimental butterflies (dashed line) resighted at Manta Real (a), Maquipucuna (b), Agua Caliente (c) and Tinalandia (d). To ensure even coverage, all possible locations to resight butterflies were visited daily on a rotating basis (except for Tinalandia where these sites were visited every 1.5 days).

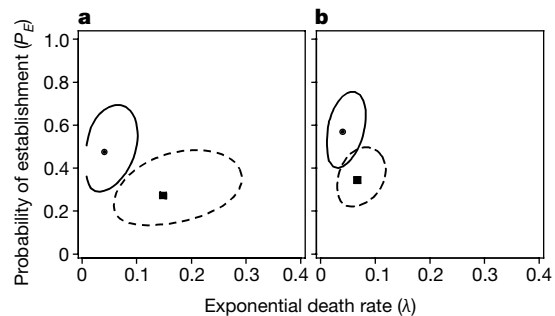


Figure 3 Probability of establishment (P_E) versus exponential death rates (λ) estimated for released butterflies at low-density sites Manta Real, Agua Caliente and Tinalandia summed (a), and at all sites combined (b). Butterflies with lower life expectancies have lower estimates of P_E and higher estimates of λ . The maximum likelihood estimates for P_E and λ for control (circles) and experimental (squares) butterflies are shown. In-likelihood profiles for these two parameters are shown by the two-unit support ellipses (control (solid line) and experimental (dotted line) butterflies) that are roughly equal to the 95% confidence limits for any one parameter¹⁹.

Table 1 Details of release experiments

Release site*	Source site	Dates	Treatments†	Release density	Total released	Total resighted
Manta Real	El Copal	15/16–28 July, 1993	C (yellow)	Low (2 per 163 m)	9	5
Manta Real	El Copal	15/16–28 July, 1993	E (white)	Low (2 per 163 m)	18	3
Maquipucuna	El Padrino	3–23 Aug, 1994	C (yellow)	High (2 per 40 m)	16	10
Maquipucuna	El Padrino	3–23 Aug, 1994	E (white)	High (2 per 40 m)	21	10
Agua Caliente	El Copal	23 Nov–5 Dec, 1994	C (yellow)	Low (2 per 164 m)	11	4
Agua Caliente	El Copal	23 Nov–5 Dec, 1994	E (white)	Low (2 per 164 m)	12	3
Tinalandia	El Copal	24 June–7 July, 1995	C (white)	Low (2 per 193 m)	14	5
Tinalandia	El Copal	24 June–7 July, 1995	E (yellow)	Low (2 per 193 m)	16	3
Overall				Control	50	24
				Experimental	67	19
				Total	117	43

* Elevations and locations of release sites: Manta Real, (400–500 m) 2° 34.73' S, 79° 20.92' W; Maquipucuna, (1,150–1,300 m) 0° 7.42' N, 78° 37.77' W; Agua Caliente, (200–350 m) 2° 37.07' S, 79° 28.99' W; Tinalandia, (600–800 m), 0° 17.88' S, 79° 3.26' W. Site elevations and locations for source sites: El Copal, (885 m) 0° 52.81' S, 79° 0.496' W and El Padrino, (775 m) 0° 0.73' S, 78° 59.21' W. † Treatments are determined by the colour of the dominant co-model species: C, controls; E, experimentals (see Methods).

species with warning coloration is a challenge to müllerian mimicry theory: distasteful butterflies should converge on a single, shared warning coloration with their müllerian co-model³. However, this experiment shows that polymorphic *H. cydno* from western Ecuador seem to be under strong geographically variable selection alternately favouring mimicry with two different co-model species. It also seems that selection against unfamiliar phenotypes may be relaxed at higher density. Thus, geographical variation in selection for mimicry coupled with weak selection against morphs when they are common⁸ explains the puzzling existence of this and other polymorphic müllerian mimics²³. Furthermore, multiple co-models coupled with relaxed selection at higher density may allow a shifting balance favouring new warning-colour patterns^{8,24–26}. This phenomenon may be particularly important not only in the origin and maintenance of new warning colours but also in the adaptive radiation of new mimetic species^{24,25,27}. □

Methods

Butterfly capture and release

I captured *H. cydno* at polymorphic source sites and transported them to release sites in glassine envelopes in a moist cooler (Table 1). I numbered, sterilized⁸ and fed the butterflies a sucrose solution at the release sites, and held them overnight in large cages. I released healthy butterflies the next day in pairs (one experimental and one control) or in triplets (two experimentals and one control, to increase the likelihood of resighting experimentals) at even intervals along release trails (Table 1). The geometry of release trails is similar at all sites (trails parallel a creek or river). I calculated the release density from the number of butterflies divided by the trail distance (m). I reduced known handling effects on recaptured butterflies²⁸ by resighting individual butterflies with 10 × 42 binoculars or a 20 × 60 spotting scope. I began resighting 12 h after release and continued until nearly all of the butterflies had disappeared from the release sites.

Treatments

At all release sites, co-models outnumber *H. cydno* (typically in a 4:1 ratio, range 2:1–36:1)⁸. At three release sites (Manta Real, Maquipucuna and Agua Caliente), abundant yellow co-models *H. eleuchia* co-occur with exclusively yellow *H. cydno*. At these sites, transferred yellow *H. cydno* butterflies are controls and white butterflies are experimentals⁸. At the fourth release site, Tinalandia, both co-model species co-occur with white and yellow morphs of *H. cydno*⁸. However, before and during the release at Tinalandia white *H. sapho* outnumbered yellow *H. eleuchia* by 5:1, thus providing the reciprocal treatment of a 'white' dominated release site⁸.

Data analysis

The daily probability of resighting a butterfly (θ_i) is estimated as a function of daily resighting effort E_i (total native butterflies of the three species encountered in a given day i , $[\theta_i = \alpha E_i]$)¹¹. I measured survival differences²¹ using maximum likelihood to estimate parameters of a simple survival model, $P_E e^{-\lambda t}$. P_E indicates the day 0 to day 1 probability of establishment or survival, and $e^{-\lambda t}$ indicates the proportion of these butterflies still alive (or present) at day t with death rate λ . Statistical analysis of survival data also takes into account θ_i (equations not shown). The complete model equations used are found in the appendix of ref. 11. Life expectancy (LE) is estimated as P_E/λ (ref. 11). Selection against the unfamiliar colour pattern ($s = 1 - LE_{exp}/LE_{con}$) where exp is experimental and con is control is directly proportional to the reduction in experimental life expectancy relative to controls¹¹.

The analysis proceeded in two steps: model parameter estimation for all combinations of treatments, release sites and different sexes (the 'full' model), followed by reduction to include only statistically significant effects^{19–21,29}. The full model fitted the observed number of butterflies resighted on a given day (goodness-of-fit test, $\chi^2_{66} = 59.13$, $P = 0.71$) and thus represents a reasonable starting point for the analysis²¹. Because the full model fit the data well I used likelihood ratio tests (LRT)^{19,21,29} and Akaike information criteria (AIC)²⁰ to reduce it to include only statistically significant main effects (and interactions), such as sex or site differences in α , P_E or λ , before testing for differences in these parameters between experimental and control butterflies²¹.

I considered first statistically significant variation in the resighting coefficient α (as suggested by ref. 21). Maximum likelihood estimates of α were higher for experimentals ($\alpha = 0.0103$) than for controls ($\alpha = 0.00798$; LRT, $G_1 = 4.59$, $P = 0.032$; AIC = 309.16, 2 parameters)^{19,20}. This indicates that experimental butterflies may stand out to human observers because they differ from butterflies with local warning colour patterns (natives and controls). Experimentals were easier to resight than controls at Maquipucuna, Agua Caliente and Tinalandia but not at Manta Real. At Manta Real, females were more likely to be resighted than males ($G_1 = 4.08$, $P = 0.0435$). To avoid potential resighting biases²¹, I estimated an intermediate resighting model that included treatment-based differences in α (control $\alpha = 0.00743$, experimental $\alpha = 0.0106$) from the three sites combined (Maquipucuna, Agua Caliente and Tinalandia, $G_1 = 5.63$, $P = 0.018$) and sex-based differences in α at Manta Real (female $\alpha = 0.013$, male $\alpha = 0.0069$). Other than effects on resighting probability at Manta Real, sex did not affect parameter estimates. I used this resighting model with the lowest AIC (306.0, four parameters) to calculate the daily probability of capture (θ_i) for the analysis of survival parameters²¹. For survival the

combined LRT and AIC analysis indicated that the best combination of P_E and λ include additive effects of site and treatment with 12 parameters (site + treatment model). Sex had no main effect on survival parameters ($G_4 = 6.60$, $P = 0.16$), and none of the interactions was statistically significant^{21,29}. I used this final model to assess the experimental effects of treatments on survival.

Programming

I programmed likelihoods using Visual Basic (Microsoft, Redmond, Washington, 1994). I located global maxima by passing ln likelihoods to the SOLVER function in Microsoft Excel (Microsoft, Redmond, Washington, 1994). All runs using different initial parameter values converged on the same maximum likelihood estimate. Visual inspections showed that all likelihood profiles were smooth and continuous with a single global maximum.

Received 14 June; accepted 10 October 2000.

- Müller, F. *Ituna* and *Thyridia*: a remarkable case of mimicry in butterflies. *Trans. Ent. Soc. Lond.* **1879**, xx–xxix (1879).
- Wickler, W. *Mimicry in Plants and Animals* (Wiedenfeld and Nicolson, London, 1968).
- Turner, J. R. G. Butterfly mimicry—the genetical evolution of an adaptation. *Evol. Biol.* **10**, 163–206 (1977).
- Fisher, R. A. *The Genetical Theory of Natural Selection* (Dover Publications, New York, 1958).
- Holmgren, N. M. & Enquist, M. Dynamics of mimicry evolution. *Biol. J. Linn. Soc.* **66**, 145–158 (1999).
- Gavrilets, S. & Hastings, A. Coevolutionary chase in two-species systems with applications to mimicry. *J. Theor. Biol.* **191**, 415–427 (1998).
- Alatalo, R. & Mappes, J. Tracking the evolution of warning signals. *Nature* **382**, 708–710 (1996).
- Kapan, D. D. *Divergent Natural Selection and Müllerian Mimicry in Polymorphic Heliconius cydno* (Lepidoptera: Nymphalidae). Thesis, Univ. of British Columbia (1998).
- Joron, M. & Mallet, J. L. Diversity in mimicry: paradox or paradigm. *Trends Ecol. Evol.* **13**, 461–466 (1998).
- Benson, W. W. Natural selection for Müllerian mimicry in *Heliconius erato* in Costa Rica. *Science* **176**, 936–939 (1972).
- Mallet, J. L. & Barton, N. H. Strong natural selection in a warning-colour hybrid zone. *Evolution* **43**, 421–431 (1989).
- Endler, J. A. Frequency-dependent predation, crypsis and aposematic coloration. *Phil. Trans. R. Soc. Lond. B* **319**, 459–472 (1988).
- Gilbert, L. E. in *Plant-Animal Interactions: Evolutionary Ecology in Tropical and Temperate Region* (eds Lewinsohn, P. W., Wilson, T. M., Fernandes, G. & Benson, W. W.) 403–427 (John Wiley and Sons, New York, 1991).
- Brower, L. P., Brower, J. V. Z. & Collins, C. T. Experimental studies of mimicry. 7. Relative palatability and Müllerian mimicry among Neotropical butterflies of the subfamily Heliconiinae. *Zoologica* **48**, 65–83 (1963).
- Chai, P. Field observations and feeding experiments on the responses of rufous-tailed jacamars (*Galbula ruficauda*) to free-flying butterflies in a tropical rainforest. *Biol. J. Linn. Soc.* **29**, 161–189 (1986).
- Brower, L. P., Cook, L. M. & Croze, H. J. Predator responses to artificial Batesian mimics released in a neotropical environment. *Evolution* **21**, 11–23 (1967).
- Cook, L. M., Brower, L. & Alcock, P. J. An attempt to verify mimetic advantage in a neotropical environment. *Evolution* **23**, 339–345 (1969).
- Waldbauer, G. P. & Sternburg, J. G. Saturniid moths as mimics: an alternative interpretation of attempts to demonstrate mimetic advantage in nature. *Evolution* **29**, 650–658 (1975).
- Edwards, A. W. F. *Likelihood* 2nd edn (Johns Hopkins Univ. Press, Baltimore, 1992).
- Akaike, H. in *International Symposium on Information Theory* 2nd edn (eds Petran, B. N. & Csáki, F.) 267–281 (Akadémiai Kiadó, Budapest, 1973).
- Lebreton, J. D., Burnham, K. P., Clobert, J. & Anderson, D. R. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecol. Monogr.* **62**, 67–118 (1992).
- Dunlap-Pianka, H., Boggs, C. L. & Gilbert, L. E. Ovarian dynamics in heliconiine butterflies: programmed senescence versus eternal youth. *Science* **197**, 487–490 (1977).
- Brown, K. & Benson, W. W. Adaptive polymorphism associated with multiple Müllerian mimicry in *Heliconius numata* (Lepid. Nymph.). *Biologica* **6**, 205–228 (1974).
- Mallet, J. L. & Joron, M. Evolution of diversity in warning colour and mimicry: Polymorphisms, shifting balance, and speciation. *Annu. Rev. Ecol. Syst.* **30**, 201–233 (2000).
- Mallet, J. L., Mallet, J. & Singer, M. C. Individual selection, kin selection, and the shifting balance in the evolution of warning colours: the evidence from butterflies. *Biol. J. Linn. Soc.* **32**, 337–350 (1987).
- Turner, J. R. & Mallet, J. L. Did forest islands drive the diversity of warningly coloured butterflies? Biotic drift and the shifting balance. *Phil. Trans. R. Soc. Lond. B* **351**, 835–845 (1996).
- Mallet, J. L., McMillan, W. & Jiggins, C. in *Endless Forms: Species and Speciation* (eds Berlocher, S. & Howard, D.) 390–403 (Oxford Univ. Press, New York, 1998).
- Mallet, J. L., Longino, J. T., Murawski, D., Murawski, A. & Gamboa, A. S. Handling effects in *Heliconius*: Where do all the butterflies go? *J. Anim. Ecol.* **56**, 377–386 (1987).
- Skalski, J. R., Hoffman, A. & Smith, S. G. in *Marked Individuals in the Study of Bird Populations* (eds Lebreton, J. D. & North, P. M.) 9–28 (Birkhäuser, Basel, 1993).
- Mallet, J. L. *et al.* Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics* **124**, 921–936 (1990).

Acknowledgements

I thank D. Schluter, L. Gilbert, M. Kirkpatrick, M. Singer, R. Dudley, U. Mueller, P. Schappert, W. O. McMillan and S. Bennett for valuable discussion and critical review of this manuscript, and H. Knechtel, C. Chapman, J. Page, S. Zaklan and K. Holston for field assistance. This research is supported by Earthwatch and its volunteer corps, by L. Gilbert and, in part, by a National Science and Engineering Research Council grant to D. Schluter.

Correspondence and requests for materials should be addressed to the author (e-mail: dkapan@rrpac.upr.clu.edu).

Identification of a receptor mediating Nogo-66 inhibition of axonal regeneration

Alyson E. Fournier*, Tadzia GrandPre* & Stephen M. Strittmatter

Department of Neurology and Section of Neurobiology, Yale University School of Medicine, P.O. Box 208018, New Haven, Connecticut 06520, USA

* These authors contributed equally to this work

Nogo has been identified as a component of the central nervous system (CNS) myelin that prevents axonal regeneration in the adult vertebrate CNS. Analysis of Nogo-A has shown that an axon-inhibiting domain of 66 amino acids is expressed at the extracellular surface and at the endoplasmic reticulum lumen of transfected cells and oligodendrocytes¹. The acidic amino terminus of Nogo-A is detected at the cytosolic face of cellular membranes¹ and may contribute to inhibition of axon regeneration at sites of oligodendrocyte injury^{2,3}. Here we show that the extracellular domain of Nogo (Nogo-66) inhibits axonal extension, but does not alter non-neuronal cell morphology. In con-

trast, a multivalent form of the N terminus of Nogo-A affects the morphology of both neurons and other cell types. Here we identify a brain-specific, leucine-rich-repeat protein with high affinity for soluble Nogo-66. Cleavage of the Nogo-66 receptor and other glycoposphatidylinositol-linked proteins from axonal surfaces renders neurons insensitive to Nogo-66. Nogo-66 receptor expression is sufficient to impart Nogo-66 axonal inhibition to unresponsive neurons. Disruption of the interaction between Nogo-66 and its receptor provides the potential for enhanced recovery after human CNS injury.

Assays of Nogo-A function have included growth-cone collapse, neurite outgrowth and fibroblast spreading with substrate-bound and soluble protein preparations¹⁻⁵. The extracellular Nogo-66 domain collapses dorsal root ganglion (DRG) axonal growth cones and inhibits neurite outgrowth¹. In assays of 3T3 fibroblast morphology, substrate-bound Nogo-66 does not inhibit spreading (Fig. 1b, c). As NI-250 preparations and full-length Nogo-A are non-permissive for 3T3 spreading^{2,5}, we considered whether different domains of Nogo might subserve this *in vitro* activity. Two pieces of evidence indicate that the N terminus of Nogo-A may contribute to inhibition of axonal regeneration: antibodies directed against this domain reduce the non-permissive qualities of CNS myelin²; and this domain alone can reduce cerebellar neurite outgrowth³.

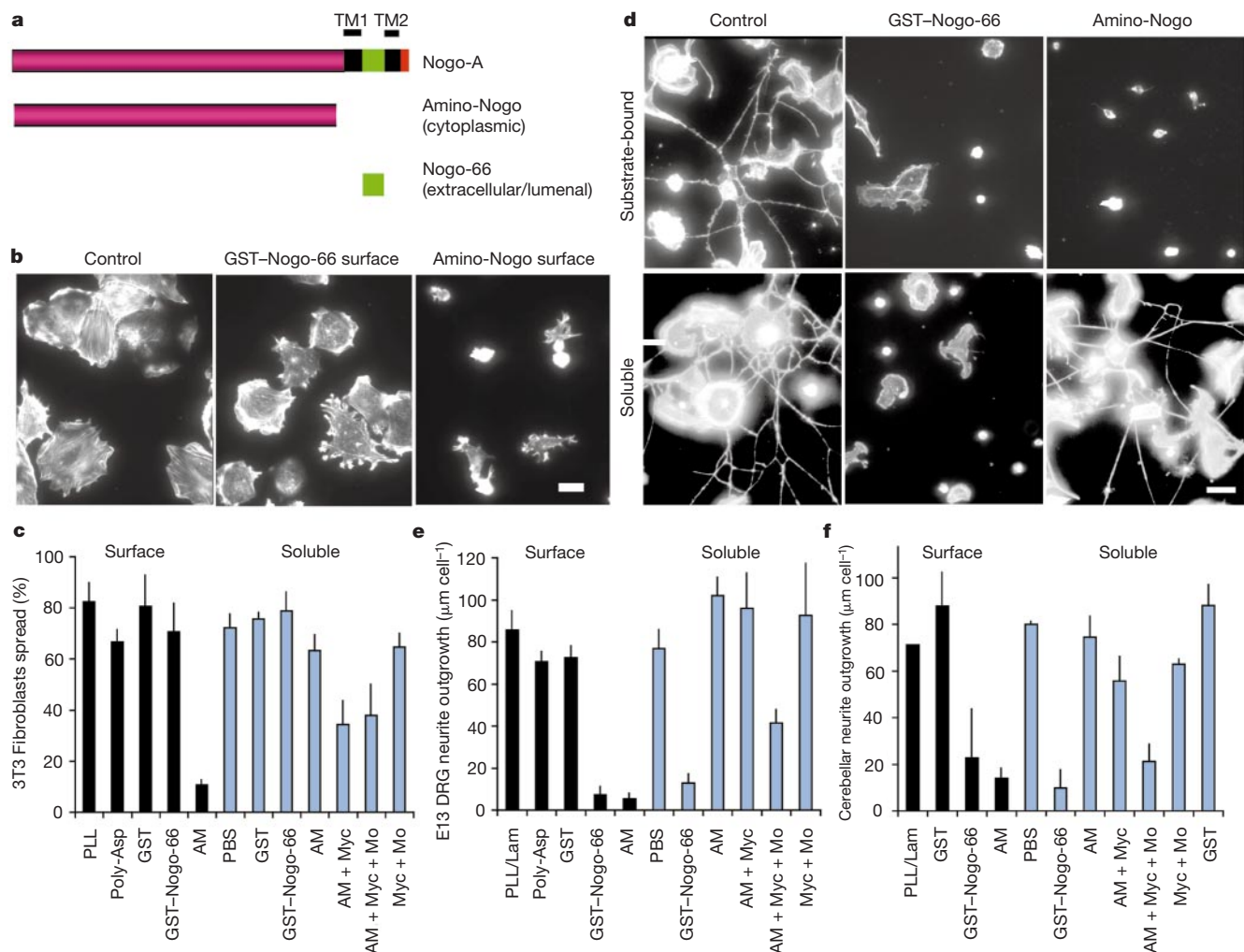


Figure 1 Comparison of Nogo domains. **a**, The Nogo proteins. **b**, NIH 3T3 fibroblasts cultured on different surfaces. Scale bar, 40 μm . **c**, 3T3 fibroblast spreading ($> 1,200 \mu\text{m}^2$) measured on Nogo-coated surfaces (black) or with soluble 100 nM Nogo preparations (blue). AM, Amino-Nogo; AM + Myc, Amino-Nogo pre-incubated with anti-Myc antibody; AM + Myc + Mo, Amino-Nogo plus anti-Myc pre-incubated with anti-

mouse IgG; Myc + Mo, anti-Myc antibody plus anti-mouse IgG antibody. **d**, Chick E12 DRGs cultured on surfaces coated with different proteins (substrate-bound) or soluble Nogo preparations (soluble). Scale bar, 40 μm . Neurite outgrowth on Nogo-coated surfaces or with soluble Nogo preparations was quantified for E13 DRG (e) or cerebellar granule cell cultures (f).

To facilitate a comparison between different Nogo-A domains, we expressed the N-terminal 1,040-amino-acid fragment (Amino-Nogo), as a MycHis-tagged protein in HEK293T cells. As predicted from immunolocalization studies¹, the protein is present in cytosolic fractions. Surfaces coated with purified Amino-Nogo protein fail to support 3T3 fibroblast spreading (Fig. 1b, c). Similar results are observed for a kidney-derived cell line, COS-7 (data not shown). The N-terminal domain therefore seems to account for the effects of full-length Nogo-A on fibroblasts.

We also exposed DRG cultures to Amino-Nogo protein (Fig. 1d, e). The fibroblast-like cells in the DRG culture do not spread on the Amino-Nogo substrate. Furthermore, axonal outgrowth is reduced to low levels on surfaces coated with Amino-Nogo. Thus, although the Nogo-66 effects are neural-specific, the inhibitory action of the Amino-Nogo domain is more generalized. When presented in soluble form at 100 nM, the Nogo-66 polypeptide collapses chick E12 DRG growth cones and nearly abolishes axonal extension¹. In marked contrast, the soluble Amino-Nogo protein seems inactive, and does not significantly modulate DRG growth-cone morphology, DRG axonal extension, or non-neuronal cell spreading (Fig. 1c–e; and data not shown). Cerebellar granule neurons have been studied previously, and soluble Amino-Nogo was presented as an Fc fusion protein—presumably in dimeric form³. We therefore considered whether these differences might explain the inactivity of soluble Amino-Nogo.

The response of mouse P4 cerebellar granule neurons and chick E13 DRG neurons to Nogo preparations is indistinguishable (Fig. 1f). Amino-Nogo dimerized with anti-Myc antibody inhibits 3T3 and COS-7 spreading (Fig. 1c; and data not shown) and tends to reduce cerebellar axon outgrowth (Fig. 1f). When further aggregated by the addition of anti-mouse IgG antibody, Amino-Nogo significantly reduces both DRG and cerebellar axon outgrowth (Fig. 1e, f). Although the Amino-Nogo protein is quite acidic, electrostatic charge alone does not account for its inhibitory effects, as poly-Asp does not alter cell spreading or axonal outgrowth (Fig. 1c, e). Both of the Nogo domains can potentially inhibit

axon growth under certain circumstances, but they possess different mechanisms of action.

On the basis of the neuronal-selective action of Nogo-66 and its high potency as a soluble monomeric ligand, we sought to identify a neuronal receptor for this Nogo domain. We chose the placental alkaline phosphatase (AP) fusion protein approach⁶. An AP–Nogo fusion protein can be purified from the conditioned medium of transfected cells in milligram amounts. This protein is biologically active as a growth-cone-collapsing agent, with an effective concentration for half-maximal response (EC_{50}) of 1 nM (Fig. 2). AP–Nogo is slightly more potent than glutathione *S*-transferase (GST)–Nogo-66, perhaps because the protein is synthesized in a eukaryotic

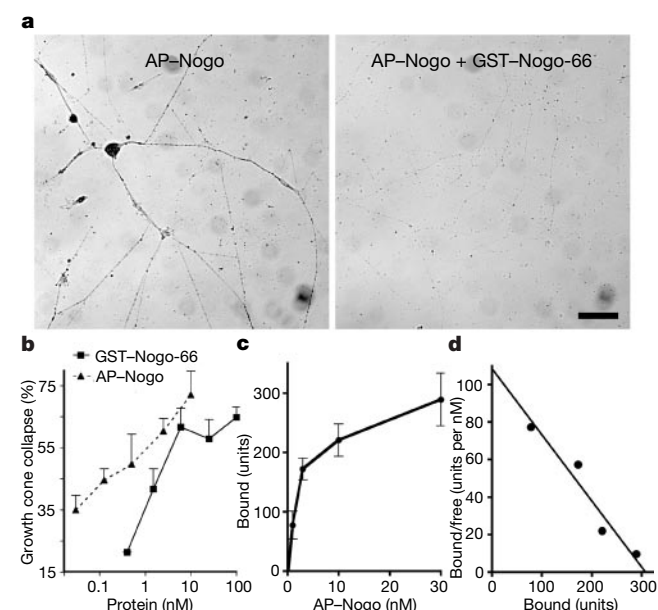


Figure 2 Ligand-binding assay for axonal Nogo-66 receptors. **a**, Dissociated chick E12 DRG neurons incubated with 10 nM AP–Nogo or 10 nM AP–Nogo + 160 nM GST–Nogo-66, and stained for bound AP. Scale bar, 40 μ m. **b**, Potency of AP–Nogo and GST–Nogo in E12 chick DRG growth-cone-collapse assays. **c**, AP–Nogo binding to DRG neurons measured as a function of AP–Nogo concentration. Error bars are from individual measurements in one of six experiments with similar results. **d**, Replotted data from **c**. K_d , 3 nM.

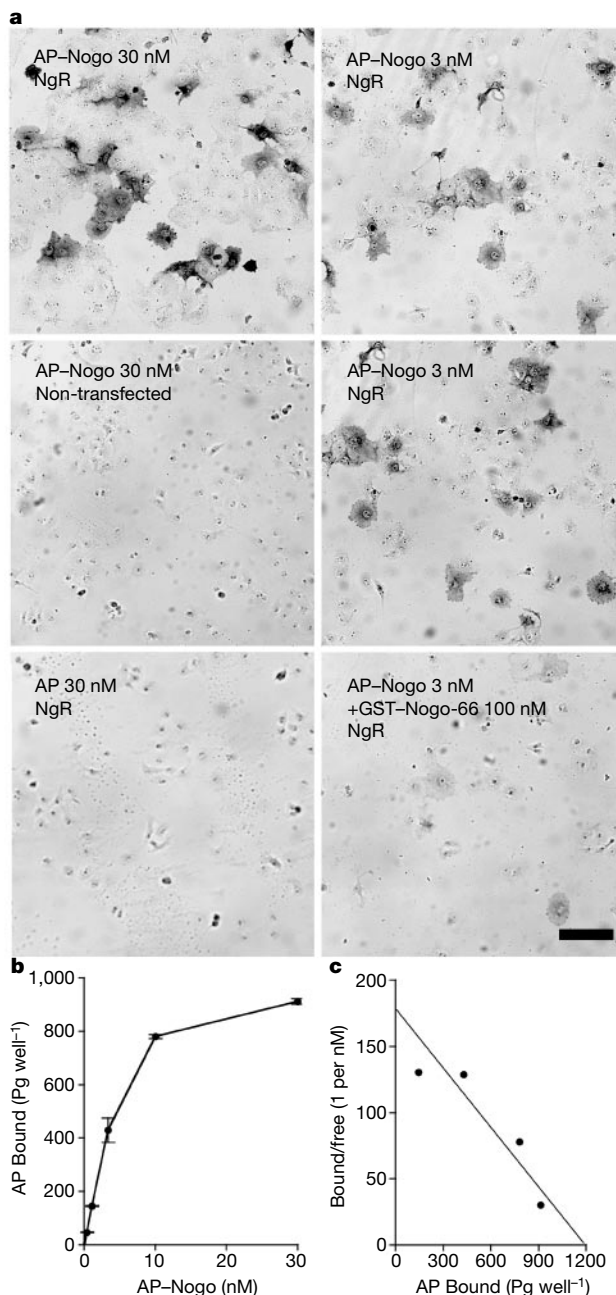


Figure 3 Nogo-66 binding to COS-7 cells expressing the Nogo-66 receptor. **a**, COS-7 cells transfected with an expression vector encoding mouse NgR or a vector control. Two days later, binding of AP–Nogo or AP was assessed. Scale bar, 200 μ m. **b**, AP–Nogo binding to NgR expressing COS-7 cells measured as a function of AP–Nogo concentration. Error bars are from individual measurements in one of six experiments with similar results. **c**, Replotted data from **b**. K_d , 7 nM.

rather than a prokaryotic cell. AP–Nogo binds to saturable, high-affinity sites on chick E12 DRG axons. Binding is blocked by excess GST–Nogo-66, consistent with competitive binding to a neuronal receptor site. As the apparent dissociation constant (K_d , 3 nM) for these sites is close to the EC_{50} of AP–Nogo in the collapse assay, the sites are probably physiologically relevant Nogo-66 receptors.

We used the AP–Nogo binding assay for expression cloning of a Nogo-66 receptor. Pools of a complementary DNA expression library from the adult brain of a mouse, representing 250,000 independent clones, were transfected into non-neuronal COS-7 cells. Non-transfected COS-7 cells do not bind AP–Nogo (Fig. 3), but transfection with two pools of 5,000 clones showed a few cells with strong AP–Nogo binding. Single cDNA clones encoding a Nogo-binding site were isolated by sib selection from each of the two positive pools. The two independently isolated clones are identical to one another except for a 100 base-pair (bp) extension of the 5' untranslated region in one clone. Transfection of these cDNAs into COS-7 cells yields a binding site with an affinity for AP–Nogo similar to that observed in E13 DRG neurons; the K_d for binding is about 7 nM (Fig. 3). AP alone does not bind with any detectable affinity to these transfected cells, indicating that the affinity is due to the 66 residues derived from Nogo. Furthermore, excess GST–Nogo-66 displaces AP–Nogo from these sites.

The cDNA directing the expression of a Nogo-binding site in COS-7 cells encodes a protein of 473 amino acids, which we have called Nogo-66 receptor (NgR). The predicted protein contains a signal sequence followed by eight leucine-rich-repeat (LRR) domains, an LRR carboxy-terminal flanking domain that is cysteine rich, a unique region and a glycosylphosphatidylinositol (GPI) anchorage site (Fig. 4). There is no full-length cDNA in GenBank with significant nucleotide sequence similarity (expectation value, $E \leq 0.05$ by BLASTN analysis); several mouse and human expressed sequence tags (ESTs) match fragments of the sequence precisely. In the LRR domains, there is moderate amino-acid sequence similarity

(up to 35% amino-acid identity) to many other proteins containing this domain. The LRR proteins sharing the greatest sequence similarity are slit1–3 and the acid-labile subunit of the insulin-like growth-factor-binding protein complex. We have identified a human homologue of the mouse NgR cDNA that shares 89% amino-acid identity. The existence of this cDNA was predicted from the mouse cDNA structure and from analysis of human genomic sequence deposited in GenBank as part of the human genome sequencing effort. The exons of the human NgR gene on chromosome 22q11 are separated by nearly 30 kilobases (kb), and the messenger RNA was not previously recognized in the genomic sequence. The predicted protein structure is consistent with a cell-surface protein capable of binding Nogo.

We next verified that the protein was associated with the cell surface through a GPI linkage. A Myc-tagged version of NgR protein with a relative molecular mass of 85,000 (M_r , 85K) is found in particulate fractions, and can be released by phosphatidylinositol-specific phospholipase C (PI-PLC) treatment, as expected for a GPI-linked protein (Fig. 4c). The GPI-linked nature of the protein suggests that there may be a second receptor subunit that spans the plasma membrane and mediates Nogo-66 signal transduction.

The simplest model for NgR expression mediating AP–Nogo binding is that the two proteins bind directly to one another. To assess their physical interaction, we incubated extracts from Myc–NgR expressing HEK293T cells with resin-bound GST–Nogo-66 or GST (Fig. 4d). NgR is selectively retained by the Nogo-containing resin, verifying that the two proteins associate as part of a protein complex.

The distribution of the mRNA expression for the NgR is consistent with a function for the protein in regulating axonal regeneration and plasticity in the adult CNS. Northern analysis shows a single band of 2.3 kb in the adult brain, indicating that the isolated NgR clone is full length (Fig. 5). Low levels of this mRNA are observed in heart and kidney but not in other peripheral tissues. In

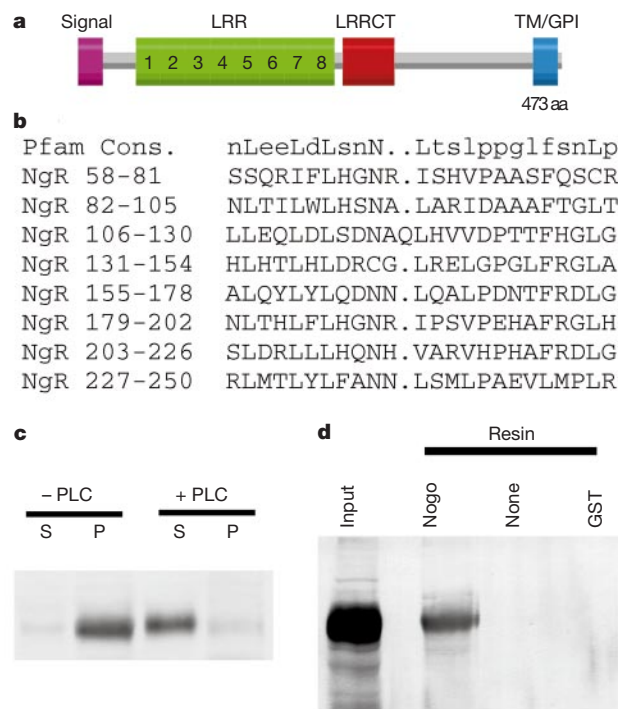


Figure 4 Structure of Nogo-66 receptor. **a**, Main structural features of the NgR protein. Signal, signal peptide; LRR, leucine-rich repeat; LRRCT, LRR C-terminal domain; TM/GPI, predicted transmembrane/glycosylphosphatidylinositol linkage. **b**, Amino-acid sequence of the LRRs is aligned with the Pfam consensus leucine-rich repeat. **c**, Extracts from Myc–

NgR-expressing cells treated with or without PI-PLC, and soluble (S) or particulate (P) fractions were analysed by anti-Myc immunoblot. **d**, Myc–NgR PI-PLC extract (Input) was incubated with glutathione-coupled GST–Nogo-66 that contained agarose (Nogo) or buffer (None), or GST (GST). Bound protein was detected by anti-Myc immunoblot.

the brain, expression is widespread and the areas richest in grey matter express the highest levels of the mRNA by northern analysis (data not shown). *In situ* hybridization shows NgR expression in cerebral cortical neurons, hippocampal neurons, cerebellar Purkinje cells and pontine neurons (Fig. 5b). NgR is expressed in those cerebral cortex pyramidal neurons whose regeneration is enhanced by IN-1 treatment⁷, and in those cerebellar Purkinje neurons whose sprouting is increased by anti-Nogo-A antibody⁸. Nogo-66 receptor mRNA is not detected in white matter (Fig. 5b), where Nogo-A is expressed by oligodendrocytes^{1,2}.

To characterize further the expression of the NgR protein, we developed an antiserum to a GST–NgR fusion protein. This antiserum detects selectively an 85K protein in NgR-expressing HEK293T cells (Fig. 5c), and specifically stains COS-7 cells expressing NgR (Fig. 5d). Immunohistological staining of chick embryonic spinal-cord cultures localizes the protein to axons, consistent with the mediation of axon-outgrowth inhibition induced by Nogo-66 (Fig. 5d). NgR expression is not found in the O4-positive oligodendrocytes (Fig. 5d) that express Nogo¹. Immunoreactive 85K protein is expressed in Nogo-66-responsive neuronal preparations from chick E13 DRGs, but to a much lesser extent in weakly responsive tissue from chick E7 DRGs and chick E7 retina (Fig. 5c). Overall, the pattern of NgR expression is consistent with the protein's mediating Nogo-66 axon inhibition.

We next investigated whether the NgR protein is necessary for

Nogo-66 action and not simply a binding site with a function unrelated to inhibition of axonal outgrowth. A first prediction is that PI-PLC treatment to remove GPI-linked proteins from the neuronal surface will render neurons insensitive to Nogo-66. This is true for chick E13 DRG neurons: PI-PLC treatment abolishes both AP–Nogo binding (data not shown) and growth-cone collapse induced by GST–Nogo-66 (Fig. 6a, c). As a control, Semaphorin 3A responses in the parallel cultures are not altered by PI-PLC treatment. However, PI-PLC treatment is expected to remove a number of proteins from the axonal surface, so this result leaves open the possibility that other GPI-linked proteins are mediating the Nogo-66 response in untreated cultures.

To show that NgR is capable of mediating Nogo-66 inhibition of axon outgrowth, we expressed the protein in neurons lacking a Nogo-66 response. We examined both DRG and retinal neurons from E7 chick embryos. The Nogo responses in the DRG neurons from this developmental stage are weak¹, but slight responses can be detected in some cultures (data not shown). E7 retinal ganglion cell growth cones are uniformly insensitive to growth-cone collapse induced by Nogo-66 (Fig. 6b, d), do not bind AP–Nogo (data not shown), and do not exhibit 85K anti-NgR immunoreactive protein (Fig. 5c). Expression of NgR in these neurons by infection with recombinant herpes simplex virus (HSV) preparations renders the axonal growth cones of the retinal ganglion cell sensitive to Nogo-66-induced collapse. Infection with a control PlexinA1-expressing

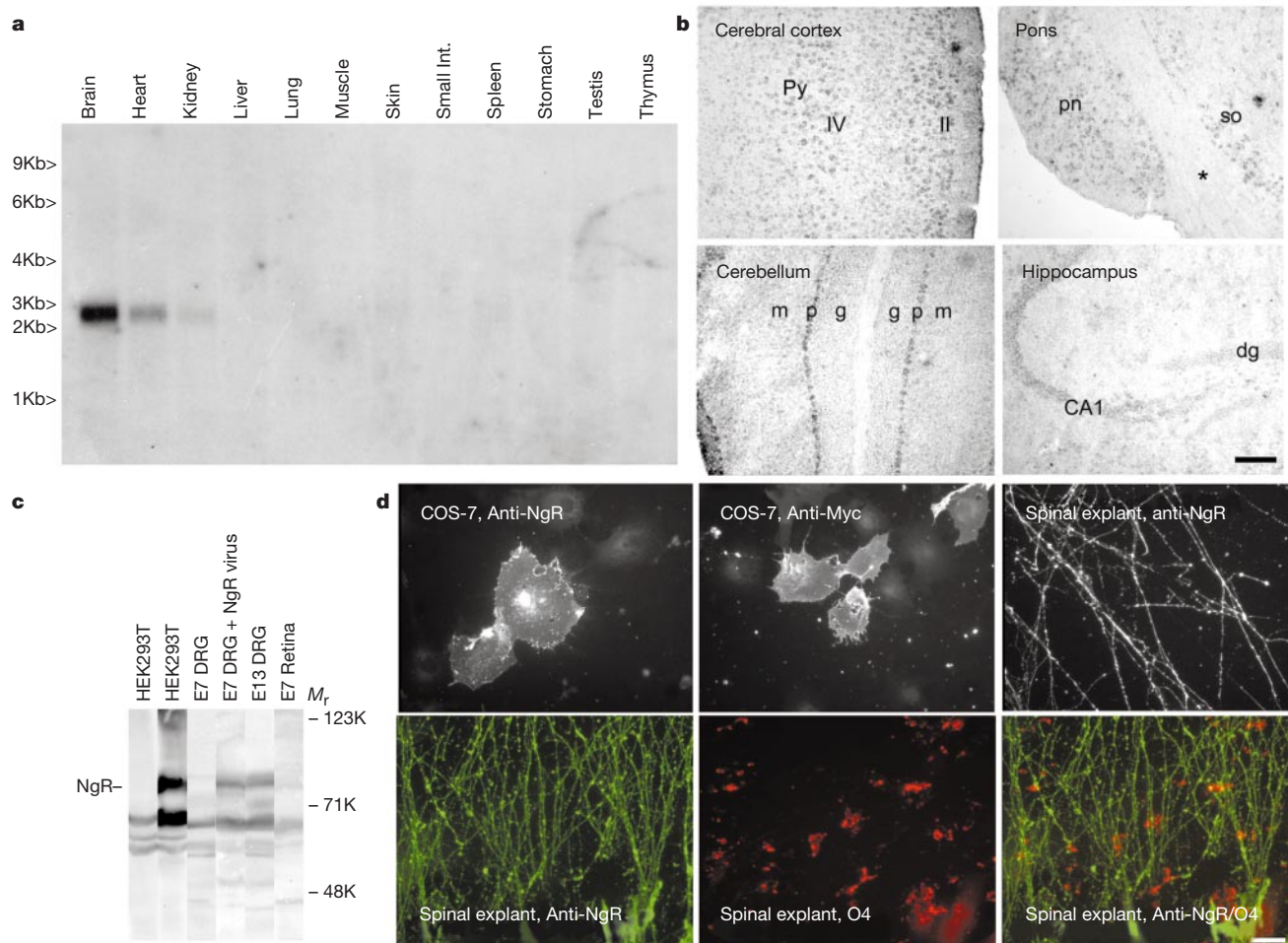


Figure 5 Distribution of Nogo-66 receptor expression. **a**, Northern analysis of NgR in mouse. The positions of RNA size markers (kb) are at the left. **b**, NgR *in situ* hybridization for adult mouse brain. II, IV, cerebral cortical layers; Py, cortical pyramidal cell layer; m, molecular, p, Purkinje cell, and g, granule cell layer of the cerebellum; dg, dentate gyrus of the hippocampus; pn, pontine nucleus of the pons; asterisk, descending pyramidal tract;

so, superior olivary nucleus. Scale bar, 500 μ m. **c**, Anti-NgR immunoblot analysis of membrane fractions from indicated cells or chick tissues. **d**, COS-7 cells expressing Myc–NgR or chick E5 spinal-cord explants (8 d *in vitro*) were stained with anti-NgR, anti-Myc or the oligodendrocyte-specific O4 antibody. Scale bars, 40 μ m (top); 80 μ m (bottom).

control HSV preparation does not alter Nogo responses. Together, these data indicate that the Nogo receptor identified here may participate in Nogo-66 inhibition of axon regeneration.

Our study identifies a GPI-linked, LRR protein as a receptor that mediates Nogo-66 inhibition of axonal outgrowth. The expression of this receptor produces a binding site for Nogo-66 with an affinity similar to that for growth-cone collapse induced by Nogo-66, and the two proteins physically associate with one another. Furthermore, expression of NgR protein is sufficient to convert axonal growth cones from a Nogo-66-insensitive to a Nogo-66-responsive state. It is clear that the extracellular 66-residue domain of Nogo and the N-terminal Nogo-A domain have different cellular effects. The biological significance of the Nogo-66 domain is supported by its neuronal specificity, its potent action as a monomeric ligand and the current identification of a high-affinity neuronal receptor for this

fragment. The NgR identified here is GPI-linked to the cell surface. This indicates that it may function as part of a receptor complex, serving as the primary Nogo-66 binding site while associated transmembrane receptor subunits are responsible for signal transduction. The Amino-Nogo domain is intracellular, has actions that are not cell-type specific, and functions only in a multivalent state. The two domains may act synergistically to inhibit outgrowth in the injured CNS. Evidence indicates that Nogo expression in the adult brain by oligodendrocytes may serve to limit axonal plasticity and regeneration after injury⁹. Therefore, the current identification of a receptor mediating Nogo-66 action should greatly facilitate the development of agents with pharmaceutical potential in a diverse group of neurological conditions, such as spinal-cord injury, brain trauma, stroke affecting white matter and chronic, progressive multiple sclerosis. □

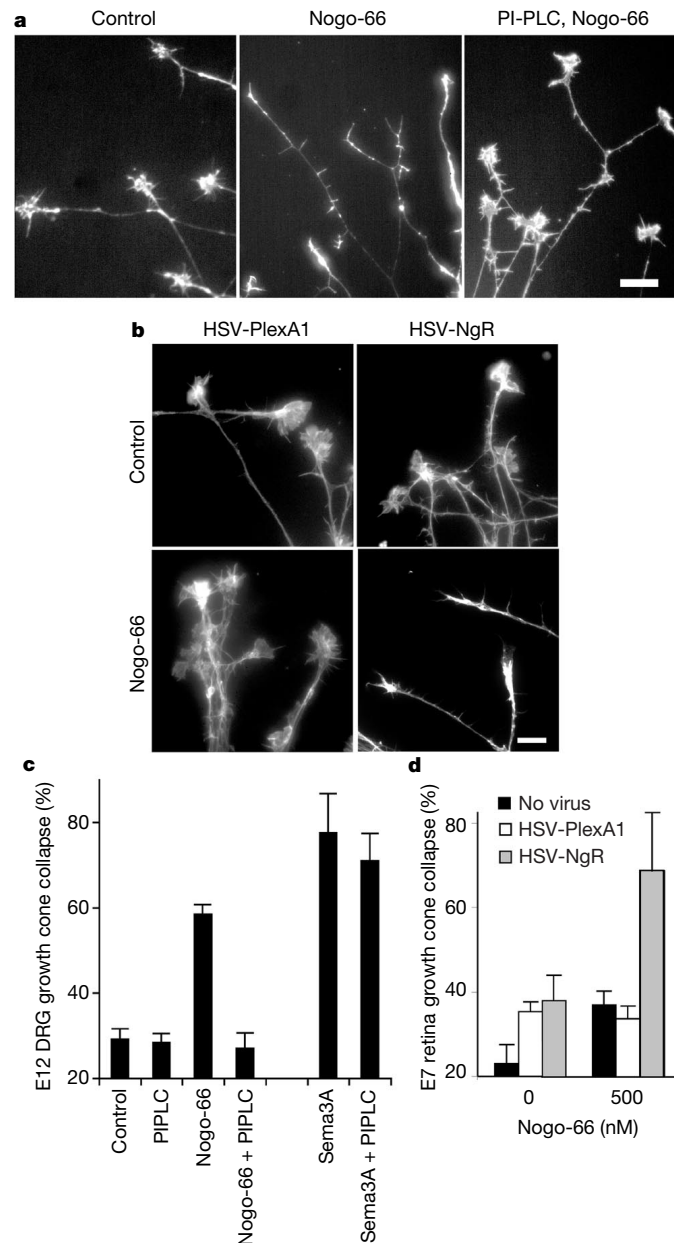


Figure 6 Nogo-66 receptor mediates growth-cone collapse by Nogo-66. **a**, Chick E12 DRG explants exposed to Nogo-66 after pretreatment with PIPLC or buffer. Scale bar, 40 μ m. **b**, E7 retinal ganglion cell explants infected with HSV-PlexA1 or HSV-Myc-NgR, and incubated with or without Nogo-66. Scale bar, 25 μ m. **c**, Growth-cone-collapse

measurements from experiments as in **a** are shown. DRG cultures treated with or without PI-PLC before exposure to 30 nM GST-Nogo-66 or 100 pM Sema3A. **d**, Quantification of growth-cone collapse in uninfected or viral-infected E7 retinal neurons as in **b**.

Methods

Nogo recombinant proteins

To express Amino-Nogo, the human Nogo-A cDNA for residues 1–1,040 was ligated into pcDNA3.1MycHis (Invitrogen, Burlingame, California) with an in-frame Myc-His tag. We transfected this plasmid into HEK293T cells and Amino-Nogo was purified with a Ni²⁺ resin¹⁰. The human Nogo-66 sequence was ligated into pcAP-5 (ref. 10) in frame with the signal sequence, 6×His tag and placental AP coding region. This plasmid was transfected into HEK293 cells, and secreted AP–Nogo was purified by Ni²⁺ affinity chromatography. GST–Nogo-66 has been described¹.

Nogo-66 receptor binding assays and expression cloning

To detect AP–Nogo binding, cultures were washed with Hanks balanced salt solution containing 20 mM sodium HEPES, pH 7.05, and 1 mg ml^{−1} bovine serum albumin (BSA) (HBH). The plates were then incubated with AP–Nogo in HBH for 2 h at 23 °C. We detected and quantified bound AP–Nogo as for AP–Sema3A¹¹. For saturation analysis of AP–Nogo bound to COS-7 cells, bound AP–Nogo protein was eluted with 1% Triton X-100. After heat inactivation of endogenous AP, we measured AP–Nogo using *p*-nitrophenyl phosphate as substrate.

For expression cloning of a Nogo-66 receptor, pools of 5,000 arrayed clones from a mouse adult-brain cDNA library (Origene Technologies, Rockville, Maryland) were transfected into COS-7 cells, and AP–Nogo binding was assessed. We isolated single NgR cDNA clones by sib selection and sequenced them. A Myc–NgR vector was created in pSecTag2-Hygro (Invitrogen) using the signal peptide of pSecTag2 fused to Myc and residues 27–473 of NgR. Human NgR cDNA was predicted from a human genomic cosmid sequence (AC007663). Oligonucleotide primers based on the predicted human cDNA amplified the cDNA from a human adult-brain cDNA library (Origene Technologies).

To assess binding of Myc–NgR to the cell membrane, particulate fractions were treated with or without 5 U PI-PLC (Sigma, St. Louis, MO) per mg of HEK293T cell protein for 1 h at 30 °C in HBH. After centrifugation at 100,000g for 1 h, we analysed equal proportions of the soluble and particulate fractions. To assess the physical interaction of NgR with Nogo-66, we incubated the PI-PLC extract (50 µg total protein) with 10 µg GST–Nogo-66 or buffer, or 10 µg GST for 1 h at 23 °C. We added glutathione-coupled agarose to bind GST and associated proteins. We analysed bound proteins by anti-Myc immunoblot.

RNA analysis

For northern blots, 1 µg poly(A)⁺ RNA from each adult mouse tissue on a nylon membrane (Origene Technologies) was hybridized with a full-length ³²P-labelled probe^{1,12}. We used digoxigenin-labelled riboprobes (nucleotides 1–1,222) and adult mouse brain sections^{1,12} for *in situ* hybridization. The sense probe produced no signal.

Nogo-66 receptor antibodies

A GST–NgR (residues 27–447) fusion protein was purified from *Escherichia coli* and used to immunize rabbits. We diluted immune serum 3,000-fold for immunoblots and 1,000-fold for immunohistology on tissue-culture samples that had been fixed by formalin. Staining of tissue was totally abolished by addition of 5 µg ml^{−1} GST–NgR.

Cell spreading, neurite outgrowth and viral infection

To measure spreading rates, subconfluent NIH 3T3 fibroblasts or COS-7 cells were plated for 1 h in serum-containing medium before fixation and staining with rhodamine-phalloidin. Glass coverslips were precoated with 100 µg ml^{−1} poly-L-lysine, washed, and then 3 µl drops of PBS containing 15 pmol Amino-Nogo, 15 pmol GST–Nogo-66, 15 pmol poly-Asp (M, 35 K, Sigma), or no protein were spotted and dried. We added soluble Nogo protein preparations (100 nM) at the time of plating. Amino-Nogo was added alone or after a pre-incubation with a twofold molar excess of anti-Myc 9E10 antibody, or with a twofold excess of anti-Myc plus a twofold excess of purified goat anti-mouse IgG.

Chick E5 spinal cord, chick E7–E13 DRG, chick E7 retina and mouse P4 cerebellar neuron culture, growth-cone-collapse assays and neurite-outgrowth assays have been described^{1,10–13}. Here, outgrowth from dissociated neurons was assessed after 12–24 h. For the substrate-bound experiments, glass chamber slides were coated with 100 µg ml^{−1} poly-L-lysine, washed, and then 3 µl drops of PBS containing 15 pmol Amino-Nogo, 15 pmol GST–Nogo-66, 15 pmol poly-Asp, or no protein were spotted and dried. After three PBS washes, we coated slides with 10 µg ml^{−1} laminin. After aspiration of laminin, dissociated neurons were added. For the soluble Nogo experiments, slides were coated with poly-L-lysine and laminin in the same fashion, and then 100 nM Amino-Nogo, 100 nM clustered Amino-Nogo, or 100 nM GST–Nogo-66 was added to the culture medium at the time of plating. After 1 d *in vitro*, some DRG explants were treated for 30 min with 1 unit ml^{−1} PI-PLC (Sigma) before the growth-cone-collapse assay.

An HSV-Myc–NgR stock was prepared as described¹⁰. We infected E7 retinal explants for 24 h. We stained some cultures infected with HSV-Myc–NgR with anti-Myc antibody to verify protein expression. Error bars reflect the s.e.m. from 4–8 determinations.

Received 3 July; accepted 30 October 2000.

- GrandPre, T., Nakamura, E., Vartanian, T. & Strittmatter, S. M. Identification of the Nogo inhibitor of axon regeneration as a reticulon protein. *Nature* **403**, 439–444 (2000).
- Chen, M. S. *et al.* Nogo-A is a myelin-associated neurite outgrowth inhibitor and an antigen for monoclonal antibody IN-1. *Nature* **403**, 434–439 (2000).

- Prinjha, R. *et al.* Inhibitor of neurite outgrowth in humans. *Nature* **403**, 383–384 (2000).
- Schwab, M. & Caroni, P. Oligodendrocytes and CNS myelin are non-permissive substrates for neurite growth and fibroblast spreading *in vitro*. *J. Neurosci.* **8**, 2381–2393 (1988).
- Caroni, P. & Schwab, M. E. Two membrane protein fractions from rat central myelin with inhibitory properties for neurite growth and fibroblast spreading. *J. Cell Biol.* **106**, 1281–1288 (1988).
- Flanagan, J. G. & Leder, P. The kit ligand: a cell surface molecule altered in steel mutant fibroblasts. *Cell* **63**, 185–194 (1990).
- Schnell, L. & Schwab, M. E. Axonal regeneration in the rat spinal cord produced by an antibody against myelin-associated neurite growth inhibitors. *Nature* **343**, 269–272 (1990).
- Buffo, A. *et al.* Application of neutralizing antibodies against NI-35/250 myelin-associated neurite growth inhibitory proteins to the adult rat cerebellum induces sprouting of uninjured Purkinje cell axons. *J. Neurosci.* **20**, 2275–2286 (2000).
- Thallmair, M. *et al.* Neurite growth inhibitors restrict plasticity and functional recovery following corticospinal tract lesions. *Nature Neurosci.* **1**, 124–131 (1998).
- Nakamura, F., Tanaka, M., Takahashi, T., Kalb, R. G. & Strittmatter, S. M. Neuropilin-1 extracellular domains mediate semaphorin D/III-induced growth cone collapse. *Neuron* **21**, 1093–1100 (1998).
- Takahashi, T., Nakamura, F., Jin, Z., Kalb, R. G. & Strittmatter, S. M. Semaphorins A and E act as antagonists of neuropilin-1 and agonists of neuropilin-2 receptors. *Nature Neurosci.* **1**, 487–493 (1998).
- Goshima, Y., Nakamura, F., Strittmatter, P. & Strittmatter, S. M. Collapsin-induced growth cone collapse mediated by an intracellular protein related to UNC-33. *Nature* **376**, 509–514 (1995).
- Huang, D. W., McKerracher, L., Braun, P. E. & David, S. A therapeutic vaccine approach to stimulate axon regeneration in the adult mammalian spinal cord. *Neuron* **24**, 639–647 (1999).

Acknowledgements

This work was supported by grants to S.M.S. from the NIH and the Christopher Reeve Paralysis Foundation. A.F. was an FCAR research fellow. S.M.S. is an Investigator of the Patrick and Catherine Weldon Donaghue Medical Research Foundation.

Correspondence and requests for materials should be addressed to S.M.S. (e-mail: stephen.strittmatter@yale.edu). The mouse and human Nogo-66 receptor mRNA sequences are deposited in GenBank under accession numbers AF283462 and AF283463, respectively.

Maize *yellow stripe1* encodes a membrane protein directly involved in Fe(III) uptake

Catherine Curie^{*†}, Zivile Panaviene^{†‡}, Clarisse Loulergue^{*},
Stephen L. Dellaporta[§], Jean-Francois Briat^{*} & Elisabeth L. Walker[‡]

^{*} *Biochimie et Physiologie Moléculaire des Plantes, Centre National de la Recherche Scientifique (Unité Mixte de Recherche 5004), Institut National de la Recherche Agronomique, Université Montpellier 2 et École Nationale Supérieure d'Agronomie, Place Viala, F-34060 Montpellier cedex 1, France*

[‡] *Biology Department, University of Massachusetts, Amherst, Massachusetts 01003, USA*

[§] *Molecular, Cellular & Developmental Biology Department, Yale University, New Haven, Connecticut 06520-8104, USA*

[†] *These authors contributed equally to this work*

Frequently, crop plants do not take up adequate amounts of iron from the soil, leading to chlorosis, poor yield and decreased nutritional quality. Extremely limited soil bioavailability of iron has led plants to evolve two distinct uptake strategies: chelation, which is used by the world's principal grain crops^{1,2}; and reduction, which is used by other plant groups^{3–5}. The chelation strategy involves extrusion of low-molecular-mass secondary amino acids (mugineic acids) known as 'phytosiderophores', which chelate sparingly soluble iron⁶. The Fe(III)-phytosiderophore complex is then taken up by an unknown transporter at the root surface^{7,8}. The maize *yellow stripe1* (*ys1*) mutant is deficient in Fe(III)-phytosiderophore uptake^{7–10}, therefore YS1 has been suggested to be the Fe(III)-phytosiderophore transporter. Here we show that *ys1* is a membrane protein that mediates iron uptake. Expression of YS1 in a yeast iron uptake mutant restores growth specifically on Fe(III)-phytosiderophore media. Under iron-deficient conditions, *ys1* messenger RNA levels increase in both roots and shoots.

Cloning of *ys1* is an important step in understanding iron uptake in grasses, and has implications for mechanisms controlling iron homeostasis in all plants.

We isolated a family that segregated ~25% of yellow striped mutants (Fig. 1a) in a random *Ac* transposon mutagenesis population¹¹. Genomic blotting using *Ac* as a probe confirmed co-segregation of an *Ac*-containing *Sal*I restriction fragment of 9.5 kilobases (kb) with the yellow striped mutant phenotype in this family (Fig. 1b). Complementation assays (data not shown) were performed with the new yellow stripe mutation to show that it is allelic to *ys1*, and so is designated *ys1-m1::Ac*. Linkage of *ys1-m1::Ac* and *pr1* on chromosome 5 was tested (see Methods), and confirmed that the new mutation is linked to *pr1*.

We prepared a genomic library from the DNA of a mutant plant, and a clone containing the co-segregating 9.5-kb *Sal*I insert was identified using *Ac* sequences as a probe (Fig. 1d). A flanking probe containing sequences adjacent to the *Ac* element (YS1-XX; see Fig. 1d) was prepared and used on a second genomic blot of the original family segregating for the yellow stripe mutation (Fig. 1b). Each mutant individual contained the 9.5-kb *Sal*I fragment, as did heterozygous wild-type plants. One mutant plant contained a 5.2-kb *Sal*I fragment, which is the expected size of a fragment after transposition of *Ac* from the 9.5-kb fragment. Notably, neither heterozygous nor homozygous wild-type plants showed the expected 5.2-kb *Sal*I fragment. The lack of the 5.2-kb fragment is probably due to cytosine methylation of the *Sal*I sites in the wild-type *Ys1* allele. To confirm the co-segregation analysis, we prepared DNA from a second family that segregated the yellow stripe mutation, so that co-segregation in a new set of individuals could be tested. The DNA was digested with *EcoRV*, an enzyme that is insensitive to methylation (Fig. 1c). On these blots, the smaller fragment (lacking *Ac*) co-segregated with the wild-type phenotype, as expected.

The YS1-XX probe was used to screen a complementary DNA library prepared from roots of iron-deficient maize plants¹². Three full-length or nearly full-length *ys1* cDNAs were recovered.

Although the precise sizes of the three cDNAs differed because of alternative polyadenylation sites and sizes of 5' untranslated regions (UTRs), they all encoded identical proteins. The sequence of these cDNAs indicates that the YS1 protein is 682 amino-acids long and contains 12 putative transmembrane domains (Fig. 2a), thus YS1 is likely to be localized to a membrane and has a structure consistent with its being a transporter protein.

The amino-acid sequence predicted from the *ys1* cDNA does not show strong similarity to any protein of known function in the various sequence databases, but it does show strong similarity to many expressed sequence tag (EST) clones in diverse plant species including both monocots and dicots, gymnosperms and mosses. The amino-acid sequence of YS1 also shows strong, full-length similarity to eight predicted *Arabidopsis* proteins (which we have designated YELLOW STRIPE1-LIKE (YSL) 1–8) of unknown function: YSL1 (491/665 residues 73% similarity; GenBank accession no. ATAC002343; PID no. AAB63613.1), YSL2 (511/658 residues 77% similarity; GenBank accession no. AB016884), YSL3 (516/668 residues 76% similarity; GenBank accession no. AB015476), YSL4 (452/644 residues 69% similarity; GenBank accession no. AB010072), YSL5 (460/680 residues 67% similarity; GenBank accession no. AB022219), YSL6 (430/604 residues 70% similarity; GenBank accession no. AB026649), YSL7 (472/674 residues 69% similarity; GenBank accession no. AC007234), and YSL8 (454/672 residues 67% similarity; GenBank accession no. AC007932; PID no. AAD49762.1).

YS1 also belongs to a gene family in maize, as there are three related maize ESTs present in GenBank. YS1 shows weak but significant similarity to a hypothetical yeast protein, YGL114 (36% similarity), belonging to the major facilitator superfamily (MFS; reviewed in ref. 13), which includes single-polypeptide secondary carriers that typically transport small solutes in response to chemiosmotic ion gradients, and to the *EspB* gene of *Myxococcus xanthus* (39% similarity). Notably, the 50 amino-terminal amino acids of YS1 contain 48% of the glutamic-acid residues of the protein (11/23). Some of these are in the sequence REKELELELER,

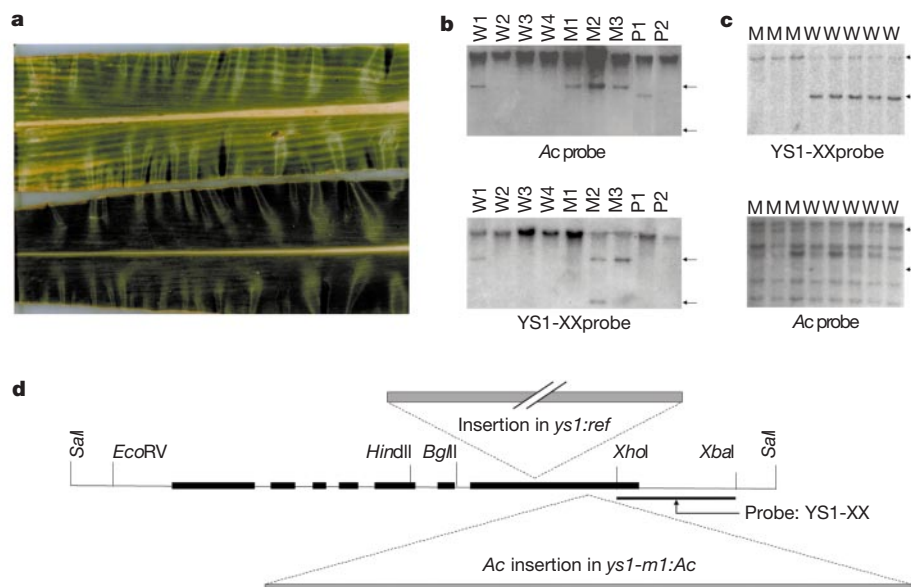


Figure 1 The *Ac*-induced yellow stripe mutation. **a**, Phenotype of homozygous mutant (above) and wild-type (below) maize leaves. **b**, Genomic blots of individuals from the original family segregating phenotypically wild-type (W1–W5) and mutant (M1–M3) individuals. DNA from the parental strains, *P-V* (P1; the *Ac*-donor locus) and *r-m3* (P2) are also shown. All samples were digested with the restriction enzyme *Sal*I. The *Ac* probe was the internal *Hind*III fragment of *Ac*. The YS1-XX probe is indicated in **d**. Arrows indicate the positions of the fragments hybridizing to the YS1-XX probe. **c**, Genomic blot

analysis of individuals from a second family segregating phenotypically wild-type (W) and mutant (M) individuals. All samples were digested with the restriction enzyme *EcoRV*. The blot was first probed with YS1-XX (above), and was then stripped and re-probed with the *Ac* probe (below). Arrows indicate the positions of the fragments hybridizing to the YS1-XX probe. **d**, Map of the *ys1* gene. Exons are indicated by black boxes. The positions of the *Ac* element in the *ys1-m1::Ac* allele and the retrotransposon element in the *ys1::ref* allele are indicated above and below. The probe fragment YS1-XX is also shown.

which is reminiscent of the REGLE sequence involved in Fe(III) transport¹⁴. This abundance of glutamic-acid residues at the amino terminus is conserved among six of the eight *Arabidopsis* YSL orthologues (data not shown).

We determined the sequence of the *ys1-m1::Ac* genomic clone λ YS31. *Ac* is inserted in the coding region at amino-acid position 649 relative to the presumed start of translation (Fig. 2b). The *Ys1* wild-type and *ys1-ref* alleles were amplified from genomic DNA using primers selected on the basis of the cDNA sequence. Genomic blot analysis (data not shown) combined with polymerase chain reaction (PCR) of the corresponding genomic region indicates that the *ys1-ref* allele has a large insertion at amino-acid position 472 relative to the start of translation (Fig. 2b). Analysis of the ends of the inserted sequence indicates that it is a long-terminal repeat retrotransposon (data not shown). Two additional *ys1* mutant alleles, *ys1:74-1924-1* and *ys1:5344*, were amplified and sequenced (Fig. 2c). The *ys1:74-1924-1* mutation corresponds to a single nucleotide insertion that causes a frameshift altering the carboxy-terminal third of the protein sequence. The *ys1:5344* allele has a slightly more complicated mutation involving a 16-base-pair (bp) deletion accompanied by a 2-bp insertion that causes a frameshift starting in the last transmembrane domain of the protein. The sequence disruption in these additional *ys1* mutant alleles provides the final confirmation that we have cloned the *ys1* gene.

Saccharomyces cerevisiae double mutant *fet3fet4* (strain DEY1453) is defective in both low- and high-affinity iron(II)

uptake systems, cannot grow on iron-limited medium³, and cannot use iron complexed with the maize phytosiderophore deoxymugineic acid (Fe-DMA) for growth¹². To investigate the function of YS1 in iron transport, we tested whether expression of a *ys1* cDNA could restore growth of the *fet3fet4* mutant on medium containing Fe-DMA as the sole iron source. The *ys1* cDNA and the *Arabidopsis* IRT1 cDNA, which encodes an *Arabidopsis thaliana* iron transporter protein capable of supporting growth of the DEY1453 strain on iron citrate³ were individually transformed into the DEY1453 strain. We then performed a differential growth test using two different sources of iron in the medium, Fe-citrate or Fe-DMA, both at a low concentration (5 μ M), to determine the substrate specificity, if any, of YS1 (Fig. 3a, b). Expression of IRT1 restored growth of *fet3fet4* when Fe-citrate was provided as sole iron source, as expected, whereas expression of YS1 did not (Fig. 3a). In the presence of 5 μ M Fe-DMA, both YS1 and IRT1 expression allowed growth of *fet3fet4* mutant, possibly owing to small amounts of residual un-chelated Fe(II) present in the medium. To clarify this, the Fe-DMA medium was supplemented with 5 μ M 4,7-biphenyl-1,10-phenanthroline-disulphonic acid (BPDS), a strong Fe(II) chelator, to remove any residual Fe(II) from the Fe-DMA medium. Addition of BPDS eliminated complementation by IRT1, without affecting complementation by YS1 (Fig. 3c). The ability of YS1 to allow growth on Fe-DMA in the presence of BPDS strongly suggests that YS1 is a transporter of phytosiderophore-bound Fe(III).

The effect of Fe starvation on *ys1* gene expression was analysed using northern blot hybridization (Fig. 4). Expression of *ys1* was detected in roots of young maize seedlings, as early as 1 day after germination (Fig. 4a, 1+). *ys1* mRNA abundance increased several-fold when plants were grown in absence of iron (Fig. 4a, 1-). The same induction was observed at 5, 7 and 10 days after germination (Fig. 4a, b), showing that steady-state levels of *ys1* mRNA are increased by iron starvation in maize roots. This result agrees well with physiological studies in which maize plants grown under iron-sufficient conditions show a low, basal level of iron uptake, and

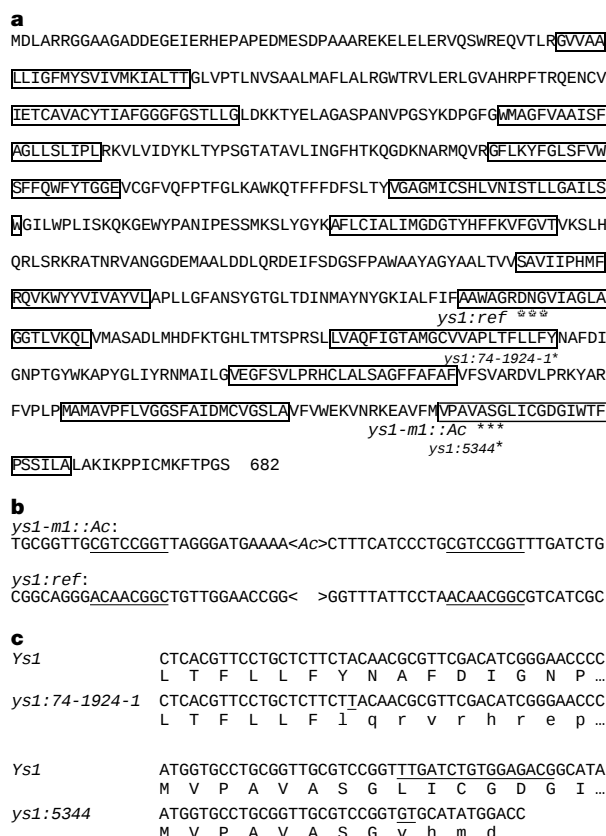


Figure 2 *ys1* sequence analysis. **a**, The predicted amino-acid sequence of YS1. The 12 putative membrane-spanning domains predicted using the SOSUI program are shown boxed²². The location of mutations in the *ys1:ref* and *ys1-m1::Ac* are shown below the sequence by asterisks and open asterisks, respectively. **b**, Genomic sequences of the *ys1:ref* and *ys1-m1::Ac* insertion sites. Target site duplications are underlined. **c**, Genomic sequences and predicted translation of non-insertion *ys1* mutations, *ys1:74-1924-1* and *ys1:5344*. Wild-type *Ys1* is shown for comparison.

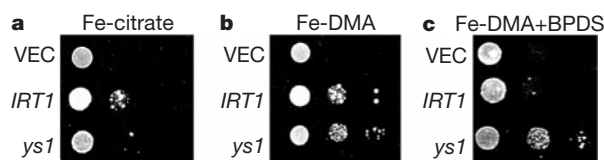


Figure 3 Growth test in the yeast iron uptake mutant. Plasmids expressing IRT1 or YS1 were introduced individually into the DEY1453 strain. The empty pYPGE15 vector was introduced as a control. Yeast growth was on minimal medium/Ura supplemented with 5 μ M Fe-citrate (**a**); 5 μ M Fe-DMA (**b**); 5 μ M Fe-DMA and 5 μ M BPDS (**c**). Three yeast cell dilutions were spotted onto the plates.

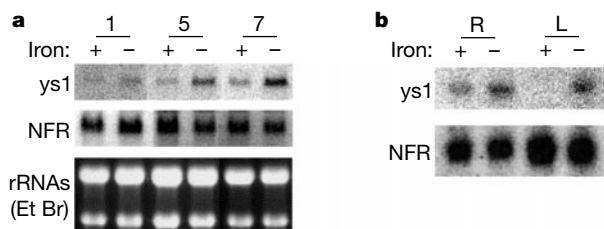


Figure 4 Regulation of *ys1* mRNA levels by iron availability in young maize plantlets. A 3' UTR *ys1* probe, obtained by PCR, was hybridized to a northern blot containing 10 μ g total RNA prepared from either roots (R) or leaves (L). **a**, RNA from roots of 1-, 5-, 7-day-old plantlets; **b**, RNA from roots and shoots of 10-day-old plantlets. The blot was stripped and hybridized to a NADPH-ferric reductase (NRF) cDNA encoding a rice cytochrome *b₅* reductase²³ as a loading control. Ethidium-bromide-stained rRNAs are also shown.

show a 2.8-fold increase in the rate of iron uptake in conditions of iron deficiency⁸.

Expression of *ys1* in leaves was investigated in 10-day-old plants grown in presence (+) or in absence (–) of iron (Fig. 4b). No *ys1* mRNA was detected in leaves of iron-sufficient plants, but a high level of accumulation was detected in leaves of iron-deficient plants. It is possible that DMA serves as an iron carrier that transports iron from cell to cell inside the plant. Indeed, DMA has been detected in leaves of rice plants¹⁵. Alternatively, nicotianamine, a Fe(II) and Fe(III) chelator structurally related to DMA¹⁶, might be a substrate for transport by YS1 in tissues other than the root. Nicotianamine is found in all plant species, not just grasses, and has been proposed to be involved in long distance Fe(II) transport in the phloem sap^{16–18}. In that regard, we note that the YSL genes of *Arabidopsis*, a species which produces nicotianamine but not mugineic acids, might have a transport role similar to that of YS1. □

Methods

Linkage analysis

The *ys1* locus is located on chromosome 5 of maize, 8 map units distal to the *pr1* locus. Heterozygous *Ys1 pr1 lys1-m1::Ac Pr1* plants were self-pollinated. Red (*pr1/pr1*) and purple (*Pr1/Pr1*) progeny were selected, and their yellow stripe phenotype was observed. The red coloured progeny were predominantly wild type with respect to yellow stripe showing that there is a clear linkage between *ys1-m1::Ac* and *pr1*, as expected. Among the purple progeny, roughly one-third of the individuals were yellow striped, again showing a clear linkage between *ys1-m1::Ac* and *pr1*.

Yeast functional complementation

Three plasmids were individually introduced into the DEY1453 (fet3fet4) strain: the *ys1* cDNA cloned in the expression vector pYPGE15, the *Arabidopsis* IRT1 cDNA cloned in the pFL61 vector¹⁹ (both expressed under the control of the strong PGK promoter¹²) and, as a control, the empty pYPGE15 vector. Minimal medium/Ura was supplemented with 5 µM Fe-citrate, 5 µM Fe-DMA, or 5 µM Fe-DMA and 5 µM BPS. Three dilutions of the culture (of optical density at 600 nm of 0.2, 0.02, and 0.002) were spotted onto the plates. Growth was carried out for 4 days at 30 °C. The Fe-DMA complex was prepared as described²⁰.

Plant growth

Plants were grown hydroponically in presence (+) or in absence (–) of iron, for 1, 5, 7 or 10 days after germination, as described²¹.

RNA extraction and northern analysis

RNA extraction and RNA blot analysis was performed as described¹². Hybridization signals were revealed after 3 days exposure, using a PhosphorImager (Storm 480, Molecular Dynamics).

Received 30 May; accepted 31 October 2000.

- Briat, J.-F. & Lobreaux, S. Iron transport and storage in plants. *Trends Plant Sci.* **2**, 187–193 (1997).
- Mori, S. Iron acquisition by plants. *Curr. Opin. Plant Biol.* **2**, 250–253 (1999).
- Eide, D., Broderius, M., Fett, J. & Gueriot, M. L. A novel iron-regulated metal transporter from plants identified by functional expression in yeast. *Proc. Natl Acad. Sci. USA* **93**, 5624–5628 (1996).
- Yi, Y. & Gueriot, M. L. Genetic evidence that induction of root Fe(III) chelate reductase activity is necessary for iron uptake under iron deficiency. *Plant J.* **10**, 835–844 (1996).
- Robinson, N. J., Procter, C. M., Connolly, E. L. & Gueriot, M. L. A ferric-chelate reductase for iron uptake from soils. *Nature* **397**, 694–697 (1999).
- Tagaki, S., Nomoto, K. & Takemoto, T. Physiological aspect of mugineic acid, a possible phyto-siderophore of graminaceous plants. *J. Plant Nutr.* **7**, 469–477 (1984).
- von Wiren, N., Mori, S., Marschner, H. & Romheld, V. Iron inefficiency in maize mutant *ys1* (*Zea mays* L. cv Yellow-Stripe) is caused by a defect in uptake of iron phytosiderophores. *Plant Physiol.* **106**, 71–77 (1994).
- von Wiren, N., Marschner, H. & Romheld, V. Uptake kinetics of iron-phytosiderophores in two maize genotypes differing in iron efficiency. *Physiol. Plant.* **93**, 611–616 (1995).
- Hopkins, B. G., Jolley, V. D. & Brown, J. C. Plant utilization of iron solubilized by oat phyto-siderophore. *J. Plant Nutr.* **15**, 1599–1612 (1992).
- Jolley, V. D. & Brown, J. C. Differential response of Fe-efficient corn and Fe-inefficient corn and oat to phytosiderophore released by Fe-efficient Coker 227 oat. *J. Plant Nutr.* **14**, 45–58 (1991).
- Dellaporta, S. L. & Moreno, M. in *The Maize Handbook* (eds Freeling, M. & Walbot, V.) 219–233 (Springer, New York, 1993).
- Loulergue, C., Lebrun, M. & Briat, J. F. Expression cloning in *Fe*²⁺ transport defective yeast of a novel maize MYC transcription factor. *Gene* **225**, 47–57 (1998).
- Pao, S. S., Paulsen, I. T. & Saier, M. H. Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.* **62**, 1–34 (1998).
- Stearman, R., Yuan, D. S., Yamaguchi-Iwai, Y., Klausner, R. D. & Dancis, A. A permease-oxidase complex involved in high affinity iron uptake in yeast. *Science* **271**, 1552–1557 (1996).

- Mori, S. *et al.* Why are young rice plants highly susceptible to iron deficiency? *Plant Soil* **130**, 143–156 (1991).
- von Wiren, N. *et al.* Nicotianamine chelates both Fe(III) and Fe(II). Implications for metal transport in plants. *Plant Physiol.* **119**, 1107–1114 (1999).
- Stephan, U. W., Schmidke, I. & Pich, A. Phloem translocation of Fe, Cu, Mn, and Zn in Ricinus seedlings in relation to the concentrations of nicotianamine, an endogenous chelator of divalent metal-ions, in different seedling parts. *Plant Soil* **165**, 181–188 (1994).
- Stephan, U. W., Schmidke, I., Stephan, V. W. & Scholz, G. The nicotianamine molecule is made-to-measure for complexation of metal micronutrients in plants. *Biomolecules* **9**, 84–90 (1996).
- Minet, M., Dufour, M. E. & Lacroute, F. Complementation of *Saccharomyces cerevisiae* auxotrophic mutants by *Arabidopsis thaliana* cDNAs. *Plant J.* **2**, 417–422 (1992).
- von Wiren, N., Gibrat, R. & Briat, J.-F. In vitro characterization of iron-phytosiderophore interaction with maize root plasma membranes: evidences for slow association kinetics. *Biochem. Biophys. Acta* **1371**, 143–155 (1998).
- Thoirion, S., Pascal, N. & Briat, J.-F. Impact of iron deficiency and iron re-supply during the early stages of vegetative development in maize (*Zea mays* L.). *Plant Cell Env.* **20**, 1051–1060 (1997).
- Hirokawa, T., Boon-Chiang, S. & Mitaku, S. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378–379 (1998).
- Bagnaresi, P. *et al.* Cloning and characterization of a maize cytochrome-*b₅* reductase with Fe³⁺-chelate reduction capability. *Biochem. J.* **338**, 499–505 (1999).

Acknowledgements

We are indebted to S. Kawai for the gift of mugineic acid and to M. L. Gueriot for the gift of IRT1 cDNA. This work was supported by a USDA NRICGP Plant Responses to the Environment Award to ELW and an NIH Award to SLD.

Correspondence and requests for materials should be addressed to E.L.W (e-mail: ewalker@bio.umass.edu). The sequence of *ys1* has been deposited at GenBank under the accession number AF186234.

CD45 is a JAK phosphatase and negatively regulates cytokine receptor signalling

Junko Irie-Sasaki*†‡, Takehiko Sasaki*†‡, Wataru Matsumoto§, Anne Opavsky||, Mary Cheng*, Grant Welstead*, Emily Griffiths*, Connie Krawczyk*, Christopher D. Richardson*, Karen Aitken||, Norman Iscove†, Gary Koretzky#, Pauline Johnson*, Peter Liul, David M. Rothstein§ & Josef M. Penninger*||

* Amgen Institute, Ontario Cancer Institute, Departments of Medical Biophysics and Immunology, University of Toronto, 620 University Avenue, Toronto, Ontario M5G 2C1, Canada

§ Department of Internal Medicine, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06529-8029, USA

|| Heart and Stroke/Lewar Centre of Excellence in Cardiovascular Research, University Health Network, University of Toronto, 200 Elizabeth Street, M5G 2C4, Canada

¶ Ontario Cancer Institute, Toronto, Ontario M5G 2M9, Canada

Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, 421 Cukic Boulevard, Philadelphia, Pennsylvania 19104, USA

* Departments of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

† These authors contributed equally to this work

The regulation of tyrosine phosphorylation and associated signalling through antigen, growth-factor and cytokine receptors is mediated by the reciprocal activities of protein tyrosine kinases and protein tyrosine phosphatases (PTPases)¹. The transmembrane PTPase CD45 is a key regulator of antigen receptor signalling in T and B cells^{2,3}. Src-family kinases have been identified as primary molecular targets for CD45 (ref. 4). However, CD45 is highly expressed in all haematopoietic lineages at all stages of

‡ Present address: Department of Pharmacology, Tokyo Metropolitan Institute of Medical Science, 3-18-22 Honkomagome, Bunkyo-ku, Tokyo, Japan 113-8613.

development⁵, indicating that CD45 could regulate other cell types and might act on additional substrates. Here we show that CD45 suppresses JAK (Janus kinase) kinases and negatively regulates cytokine receptor signalling. Targeted disruption of the *cd45* gene leads to enhanced cytokine and interferon-receptor-mediated activation of JAKs and STAT (signal transducer and activators of transcription) proteins. *In vitro*, CD45 directly dephosphorylates and binds to JAKs. Functionally, CD45 negatively regulates interleukin-3-mediated cellular proliferation, erythropoietin-dependent haematopoiesis and antiviral responses *in vitro* and *in vivo*. Our data identify an unexpected and novel function for CD45 as a haematopoietic JAK phosphatase that negatively regulates cytokine receptor signalling.

To investigate a possible involvement of CD45 in cytokine receptor signalling, we generated interleukin-3 (IL-3)-dependent bone-marrow-derived mast cell (BMMC) lines from *cd45*^{+/+} mice and mice that carry a mutation in exon 6 of the *cd45* gene (*cd45*^{-/-})² (see Fig. 1 in Supplementary Information). BMMCs from both genotypes showed similar expression levels of FcεR1, c-Kit and IL-3 receptor β-chain, indicating that a lack of CD45 does not prevent the emergence and differentiation of BMMCs. During the generation of mast cells with IL-3, we noted that the expansion rate of *cd45*^{-/-} BMMCs was greater than that of *cd45*^{+/+} BMMCs (see Fig. 1 in Supplementary Information). Proliferation of *cd45*^{-/-} BMMCs was markedly increased in response to stimulation with IL-3 over that of wild-type cells (Fig. 1a). Thus, in contrast to antigen receptor stimulation, in which a loss of CD45 leads to decreased cell growth^{2,3}, our data show that a loss of CD45 expression leads to an increased proliferation of BMMCs in response to the cytokine IL-3.

Stimulation with IL-3 triggers multiple signalling pathways such as the activation of extracellular signal-regulated kinases (ERKs), protein kinase B (PKB)/Akt and JAKs⁶. To assess the importance of CD45 in regulation of these signalling molecules, we analysed the phosphorylation of these molecules in response to IL-3. Phosphorylation of ERK1/ERK2 (Fig. 1b) and phosphatidylinositol-3-OH kinase (PI(3)K)-dependent activation of PKB/Akt (see Fig. 1 in Supplementary Information) were comparable between wild-type and *cd45*^{-/-} BMMCs. In *cd45*^{+/+} BMMCs, JAK2 tyrosine phosphorylation in response to IL-3 was readily detected. Intriguingly, IL-3-induced JAK2 tyrosine phosphorylation was significantly enhanced in *cd45*^{-/-} BMMCs (Fig. 1b). Tyrosine phosphorylation of JAK2 was also enhanced in BMMCs derived from a line of independently derived CD45-null mice³ (not shown). Consistent with JAK2 hyperphosphorylation was the observation

that JAK2 kinase activity was increased in the absence of CD45 after stimulation with IL-3 (Fig. 1c). Moreover, IL-3-induced tyrosine phosphorylation of the JAK substrates STAT3 (Tyr 705) and STAT5 (Tyr 694) in response to IL-3 was significantly increased in *cd45*^{-/-} BMMCs in comparison with wild-type BMMCs (Fig. 2a). Phosphorylation of STAT3 on Ser 727 in response to IL-3 was comparable in *cd45*^{+/+} and *cd45*^{-/-} BMMCs (Fig. 2a), indicating that the PTPase CD45 does not regulate the serine phosphorylation of STAT3. Phosphorylation of Tyr 705 in STAT3 and Tyr 694 in STAT5 is required for the dimerization and transcriptional activity of these STATs⁷. Consistent with the enhanced phosphorylation of STAT3, was the observation that stimulation with IL-3 induced more DNA-binding activity of STAT3 in BMMCs lacking CD45 than in control *cd45*^{+/+} BMMCs (Fig. 2b). Furthermore, the induction of *cyclin D1*, which mediates IL-3-dependent proliferation through STAT5 (ref. 8), was enhanced in *cd45*^{-/-} BMMCs (Fig. 2c). These results indicate that CD45 negatively regulates the IL-3-triggered JAK-STAT signalling pathway.

Because Src-family kinases are direct substrates for CD45 (ref. 4), we thought that deregulated Src-family kinase activity leads to hyperactivation of the JAK-STAT pathway in *cd45*^{-/-} BMMCs. However, inhibition of Src-family kinases with the Src-family kinase inhibitor PP2 (ref. 9) did not affect JAK2 activation in *cd45*^{+/+} or *cd45*^{-/-} BMMCs. Moreover, in IL-3-stimulated *cd45*^{+/+} or *cd45*^{-/-} BMMCs, treatment with PP2 had no effect on the increased phosphorylation of the specific residues Tyr 1007 and Tyr 1008 in JAK2, Tyr 705 in STAT3, and Tyr 1022 and Tyr 1023 in JAK1 (see Fig. 2 in Supplementary Information). Thus, Src-family kinases do not account for increased JAK-STAT activation in IL-3-stimulated mast cells. To evaluate whether CD45 could directly dephosphorylate JAKs and/or STATs, we used phosphatase assays *in vitro* with recombinant CD45 (rCD45). The rCD45 employed contained the intracellular dual PTPase domains with a relative molecular mass of 97,000 (*M_r* 97 K) (the catalytic D1 and the non-catalytic D2 domains) of CD45 but lacked its extracellular regions. As expected, rCD45 dephosphorylated Lyn in a dose-dependent manner (Fig. 3a). In contrast, tyrosine dephosphorylation of STAT3 (Fig. 3a, b) and STAT5 (not shown) did not occur after incubation with rCD45, even at high concentrations of rCD45. Interestingly, rCD45 could directly dephosphorylate JAK2 *in vitro* (Fig. 3a). In contrast, the human non-receptor tyrosine PTPase TC-PTP and the bacterial lambda PTPase did not dephosphorylate JAK2 *in vitro* (see Fig. 2 in Supplementary Information). Incubation of phosphorylated JAK2 with rCD45 in the presence of the PTPase inhibitor vanadate blocked JAK2 dephosphorylation (Fig. 3b). In addition,

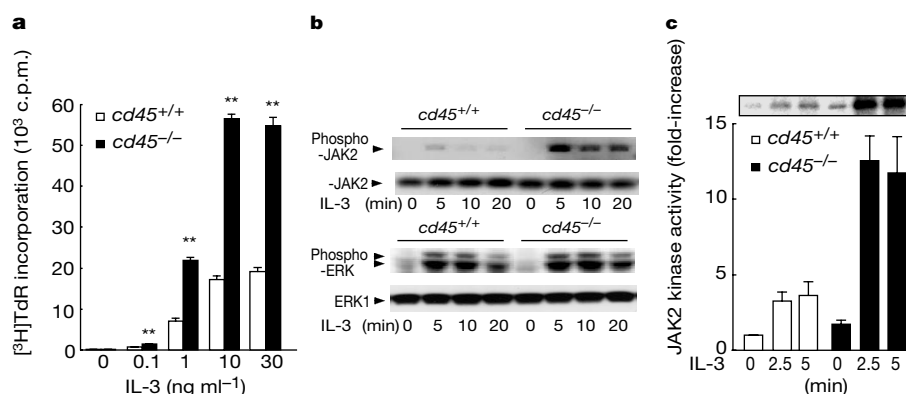


Figure 1 CD45 negatively regulates cytokine-induced activation of mast cells and JAKs. **a**, Increased proliferation of *cd45*^{-/-} BMMCs. *cd45*^{+/+} and *cd45*^{-/-} BMMCs were incubated with increasing doses of IL-3 and proliferation was assessed by [³H]thymidine incorporation. Asterisks, *P* < 0.01 between *cd45*^{+/+} and *cd45*^{-/-} BMMCs. **b**, Enhanced JAK2 phosphorylation. BMMCs were stimulated with IL-3 (30 ng ml⁻¹). JAK2 was

immunoprecipitated; this was followed by immunoblotting with anti-phosphotyrosine and anti-JAK2. ERK activation was assayed with an anti-phospho-ERK antibody. **c**, Increased JAK2 kinase activity in *cd45*^{-/-} BMMCs stimulated with IL-3 (30 ng ml⁻¹). JAK2 activity was assessed by kinase assay *in vitro*. Values (±s.e.m.) are expressed as fold increase over control (wild type, no IL3).

PTPase-dead rCD45 (D1:C828S mutant)¹⁰ did not dephosphorylate JAK2 (not shown), demonstrating that CD45 PTPase activity is responsible for JAK2 dephosphorylation.

Using site-specific and phospho-specific antibodies, we found that the critical tyrosine residues Tyr 1007 and Tyr 1008 of JAK2 are dephosphorylated by rCD45 (Fig. 3b), resulting in decreased JAK2 kinase activity (Fig. 3c). As with JAK2, rCD45 dephosphorylated JAK1 and Tyk2 on tyrosine residues *in vitro*, including the critical Tyr 1022 and Tyr 1023 residues of JAK1 (Fig. 3d). Although we have not identified all the dephosphorylation sites of JAKs by CD45, these results show that CD45 dephosphorylates functionally important tyrosine residues. It should be noted that, as with our phosphatase assays *in vitro*, Tyr 1022 and Tyr 1023 of JAK1, Tyr 1007 and Tyr 1008 of JAK2, and Tyr 1054 and Tyr 1055 of Tyk2 are indeed hyperphosphorylated in *cd45*-deficient cells (Fig. 4a; see Fig. 2 in Supplementary Information). In addition, we observed a physical association between JAK2 and the intracellular portion of a PTPase-inactive CD45 trap mutant (glutathione S-transferase (GST)–CD45 D1:C828S). Of note, the second ‘pseudo-PTPase’ domain of CD45 (GST–D2) could itself bind to JAK2 (Fig. 3e). To extend our findings *in vitro* to cells and to examine whether the PTPase activity of CD45 is required for the negative regulation of JAKs *in vivo*, we expressed wild-type and PTPase-dead CD45 (D1:C828S)¹¹ in CD45-negative COS cells (Fig. 3f). Stimulation of COS cells with

interferon- α (IFN- α) triggers the tyrosine phosphorylation of JAK1 at Tyr 1022 and Tyr 1023. Ectopic expression of wild-type CD45 resulted in decreased JAK1 tyrosine phosphorylation (Fig. 3f). Importantly, expression of the PTPase-inactive point mutant of CD45 had no effect on JAK1 tyrosine phosphorylation in response to IFN- α . Thus, CD45 can associate with JAK2 and directly regulate the tyrosine phosphorylation of JAK-family kinases.

A wide variety of cytokines are known to activate JAKs^{7,12}. The direct dephosphorylation of JAKs by CD45 raised the possibility that, in addition to IL-3 in mast cells, IFN- α -induced activation of JAK1 (Fig. 4a) and Tyk2 (see Fig. 3 in Supplementary Information) was increased in *cd45*-deficient (CD45-AS) human Jurkat T cells in comparison with the parental wild-type Jurkat cells (Fig. 4a). Re-expression of wild-type CD45 in the CD45-AS cells decreased the levels of JAK1 phosphorylation to that of the parental Jurkat cells. Importantly, re-expression of PTPase-dead CD45 (D1:C828S) in CD45-AS cells did not alter increased JAK1 phosphorylation (Fig. 4a). Thus, as in the COS cell experiments, CD45-PTPase activity is required for the negative regulation of JAK1 phosphorylation in IFN- α -activated Jurkat cells. In freshly isolated thymocytes, stimulation with IFN- α resulted in JAK1 phosphorylation that was higher in *cd45*^{-/-} cells than in wild-type thymocytes (Fig. 4b). Moreover, JAK1 and JAK3 tyrosine phosphorylation was upregulated in primary *cd45*^{-/-} B cells in response to the B-

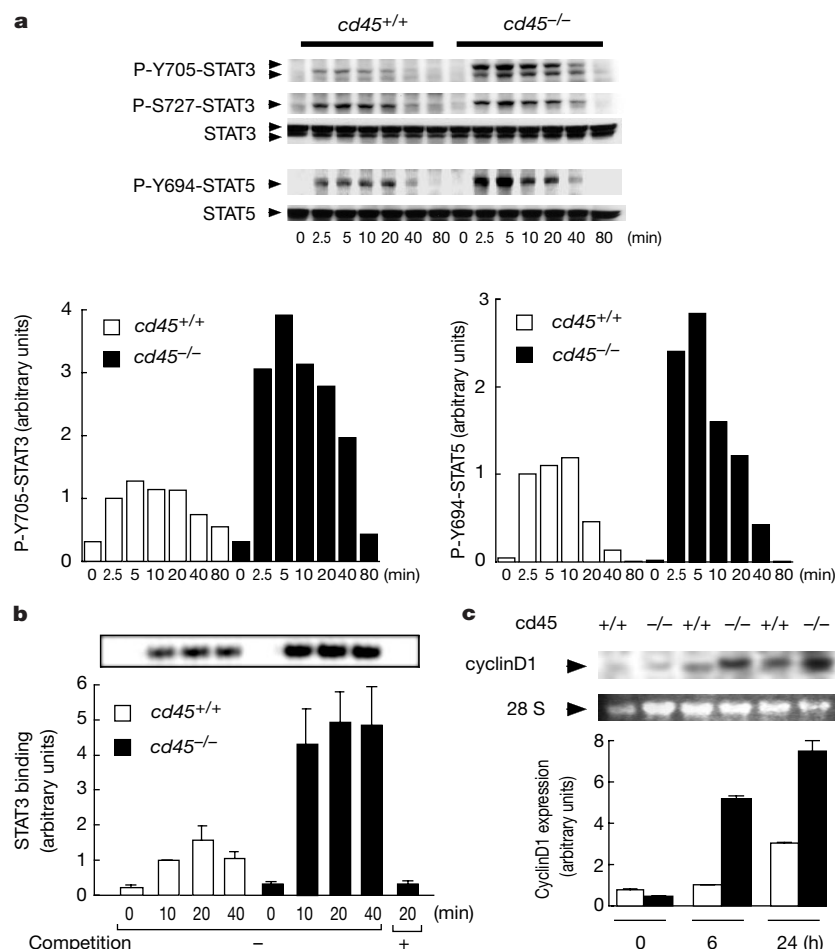


Figure 2 CD45 deficiency leads to increased tyrosine phosphorylation of STATs. **a**, Increased STAT phosphorylation in the absence of CD45. *cd45*^{+/+} and *cd45*^{-/-} BMMCs were treated with IL-3 (30 ng ml⁻¹) and phosphorylation was determined with phosphospecific antibodies against STAT3 (Tyr 705) (P-Y705-STAT3), STAT3 (Ser 727) (P-S727-STAT3) or STAT5 (Tyr 694) (P-Y694-STAT5). **b**, Increased DNA-binding activity of STAT3 in *cd45*^{-/-} BMMCs treated with 30 ng ml⁻¹ IL-3. Nuclear extracts and a

radiolabelled STAT3-specific probe were used in EMSA. In the last lane, an excess of the unlabelled DNA probe was added (competition +). The lower panel shows mean STAT3 binding ($\pm$ s.e.m.) from triplicate experiments. **c**, Enhanced induction of *cyclin D1* mRNA in the absence of CD45. Induction of *cyclin D1* was analysed by northern blot at the indicated time points after stimulation with IL-3. 28 S rRNA is shown as control.

cell cytokine IL-4 (Fig. 4c). Finally, in freshly isolated macrophages from *cd45*^{-/-} mice, Tyk2, STAT1 (Tyr 701) and STAT3 (Tyr 705), but not ERK1/2, were hyperphosphorylated in response to IFN- α (see Fig. 4 in Supplementary Information). It should be noted that our results are in contrast to previous experiments with Jurkat cells and a high dose of IFN- α that suggested that CD45 is a positive regulator of JAK-STAT signalling¹³. Our data now demonstrate that CD45 is a negative regulator of JAKs in multiple haematopoietic cell types and in response to different cytokines and antiviral IFN- α .

Cytokine-mediated activation of JAKs is important for a wide

array of cellular functions such as proliferation, differentiation, survival and host resistance to pathogens⁷. We first investigated the effect of CD45 deficiency on cytokine-dependent erythropoiesis and myelopoiesis, both of which are regulated by the JAK-STAT pathway^{14,15}. The addition of erythropoietin (EPO) to bone-marrow progenitors induces the growth and differentiation of erythroid colonies (BFU-Es; burst-forming units-erythroid) in a dose-dependent manner. The numbers of EPO-dependent erythroid colonies that differentiated from *cd45*^{-/-} progenitors were significantly increased in comparison with the numbers derived from *cd45*^{+/+}

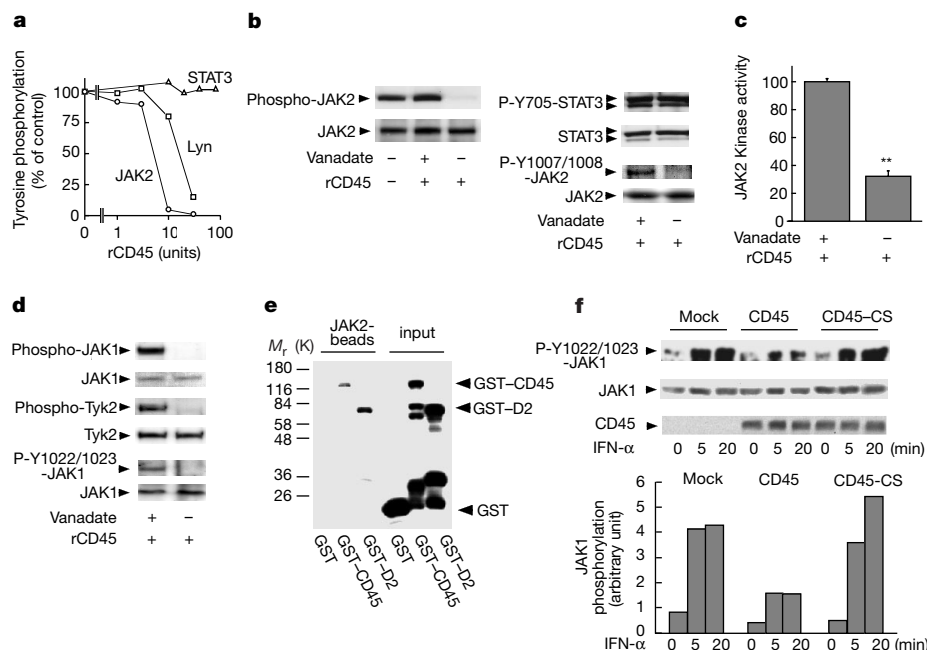


Figure 3 CD45 dephosphorylates JAKs *in vitro*. **a**, Recombinant CD45 dephosphorylates Lyn and JAK2, but not STAT3, *in vitro*. **b**, Vanadate inhibits rCD45-mediated JAK2 dephosphorylation. Phosphorylation was examined with anti-phosphoryrosine, anti-phospho-Tyr 705 STAT3 (P-Y705-STAT3), and anti-phospho-Tyr 1007/Tyr 1008 JAK2 (P-Y1007/1008-JAK2). Assays were performed with 10 U (for JAK2) or 90 U (for STAT3) of rCD45. **c**, Decreased kinase activity of JAK2 dephosphorylated by rCD45 (10 U). Asterisks, $P < 0.001$, kinase activity between non-treated and treated with CD45. **d**, CD45 dephosphorylates JAK1 and Tyk2. **e**, CD45 binds to JAK2. JAK2 was incubated

with GST alone, GST-CD45 (cytoplasmic portion of a PTPase-dead mutant), and GST-D2 (non-catalytic D2 domain of CD45). Proteins bound to JAK2-beads were detected by anti-GST. **f**, Heterotopic expression of CD45 suppresses tyrosine phosphorylation of JAK1. COS cells were transfected with wild-type CD45 (CD45), phosphatase-inactive CD45 (CD45-CS), or mock-transfected controls and stimulated at 48 h with IFN- α (500 U ml⁻¹). Phospho Tyr 1022/Tyr 1023 levels of JAK1, total JAK1 and expression levels of transfected CD45 are shown.

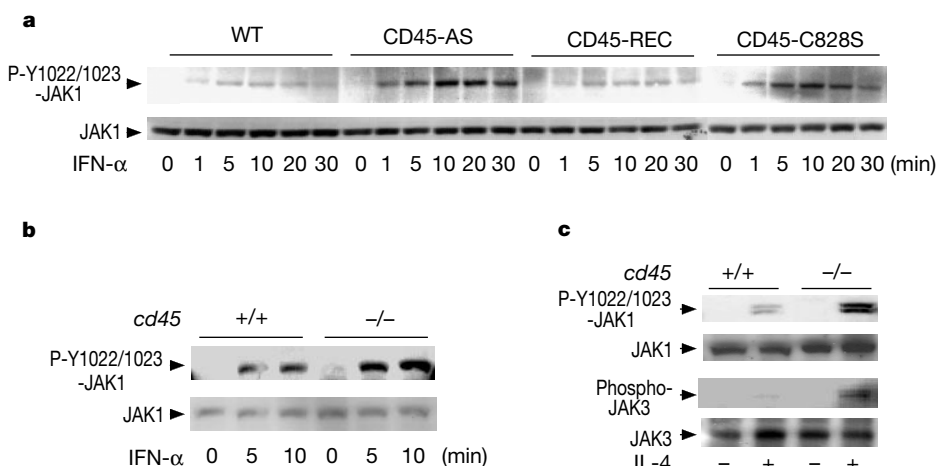


Figure 4 JAKs are hyperphosphorylated in CD45-deficient B cells, thymocytes and Jurkat T cells. **a**, Parental wild-type Jurkat cells (WT), *cd45*-deficient Jurkat cells (CD45-AS), CD45-AS cells re-expressing wild-type CD45 (CD45-REC) and CD45-AS cells re-expressing PTPase-dead CD45 (CD45-C828S) were stimulated with IFN- α (100 U ml⁻¹).

b, Freshly isolated thymocytes from *cd45*^{+/+} and *cd45*^{-/-} mice were stimulated with IFN- α (20,000 U ml⁻¹). **c**, Freshly isolated splenic B cells from *cd45*^{+/+} and *cd45*^{-/-} mice were stimulated with recombinant IL-4 (100 ng ml⁻¹) for 10 min. Tyrosine phosphorylation of JAK1 and JAK3 was monitored as described in Methods.

progenitor cells (Fig. 5a). Individual *cd45*^{-/-} erythroid colonies were also substantially greater in size and density than EPO-dependent *cd45*^{+/-} erythroid colonies. In addition, the formation of IL-3-induced mixed neutrophil/macrophage myeloid colonies was significantly increased when bone-marrow progenitors were isolated from *cd45*^{-/-} mice as opposed to *cd45*^{+/-} littermates (*cd45*^{+/-}, 139 ± 20; *cd45*^{-/-}, 206 ± 29 (*P* < 0.01); 0.1 ng ml⁻¹ IL-3, determined on day 7). Thus, a loss of CD45 in haematopoietic progenitor cells leads to increased cytokine-dependent erythropoiesis and myelopoiesis.

IFN-α and IFN-α receptor (IFN-αR) activation regulate susceptibility to viral infections *in vivo* and this effect is mediated by the JAK-STAT pathway¹⁶. IFN-α and IFN-αR-regulated signalling are critical for replication of the picornavirus Coxsackievirus B3 (CVB3) *in vitro* and CVB3 infections *in vivo*¹⁷. To test whether CD45 can regulate CVB3 replication directly *in vitro*, we measured CVB3 titres after infection of parental or *cd45*-deficient Jurkat cell lines. IFN-α suppressed viral amplification more efficiently in *cd45*-deficient cells than in parental Jurkat cells; this suppression was dependent on a functional CD45-PTPase domain (Fig. 5b). To test whether reduced CVB3 replication in *cd45*-deficient Jurkat cell lines *in vitro* would translate into altered disease pathogenesis of CVB3 infections *in vivo*, we inoculated *cd45*^{+/-} and *cd45*^{-/-} littermate mice with CVB3. Approximately 50% of *cd45*^{+/-} mice succumbed to CVB3 infections within 7 days after infection owing to the acute cytopathic effects of the virus leading to encephalitis, pancreatitis, myocarditis and hepatitis. In contrast, *cd45*^{-/-} littermates were completely protected from lethal CVB3 infections (Fig. 5c) and did not show any histological lesions of acute or chronic inflammation in the heart, pancreas, liver or brain (Fig. 5d, and not shown). Thus, the loss of CD45 renders mice resistant to lethal infections with CVB3. The role of CD45 in other viral infections *in vivo* needs to be determined.

The JAK family contains four members, JAK1, JAK2, JAK3 and Tyk2, that are differentially activated in response to different cytokines⁷. JAKs and JAK-mediated STAT activation have crucial roles in proliferation, differentiation, survival, host resistance to pathogens, and malignancies¹⁸. Thus, multiple inhibitory mechanisms and molecular backup systems must exist to regulate cytokine-triggered cellular activation. SOCS (suppressor of cytokine signalling) family proteins bind to phosphotyrosine residues within the activation loop of the JAK proteins to inhibit their kinase activity¹⁹. Moreover, the PTPase SHP1 is recruited to certain cytokine receptors and binds to JAKs²⁰. SHP1 can function as a negative regulator of JAK-STAT signalling, although it remains to be shown whether the inhibitory effects of this phosphatase is mediated by dephosphorylating signalling molecules regulating JAKs or by direct dephosphorylation of JAKs¹. We provide evidence that CD45 is a novel negative regulator of cytokine-receptor-mediated proliferation and haematopoiesis and IFN-regulated antiviral responses. Genetic inactivation of CD45 leads to hyperactivation of the JAK-STAT pathway *in vivo*. In humans, loss of CD45 expression has been reported in more than 10% of patients with acute lymphoblastic leukaemia²¹, and CD45 expression is frequently lost in Hodgkin's lymphoma cells²². Moreover, a genetic mutation of CD45 has been reported in a child that ultimately succumbed to a B-lymphoma²³.

JAK expression is not restricted to haematopoietic cells, and JAK-STAT pathways have also been implicated in the homeostasis of cardiac myocytes, mammary gland involution, and responses to insulin, growth hormone or leptin²⁴. Presumably other JAK PTPases exist in addition to the haematopoietic-restricted CD45 that regulate JAKs in non-haematopoietic cells. For example, the transmembrane PTPase LAR, a close structural homologue of CD45, has been shown to downregulate insulin receptor signalling *in vivo*²⁵. In our preliminary studies *in vitro*, LAR was indeed capable of dephosphorylating JAK2. It remains to be established by genetic means

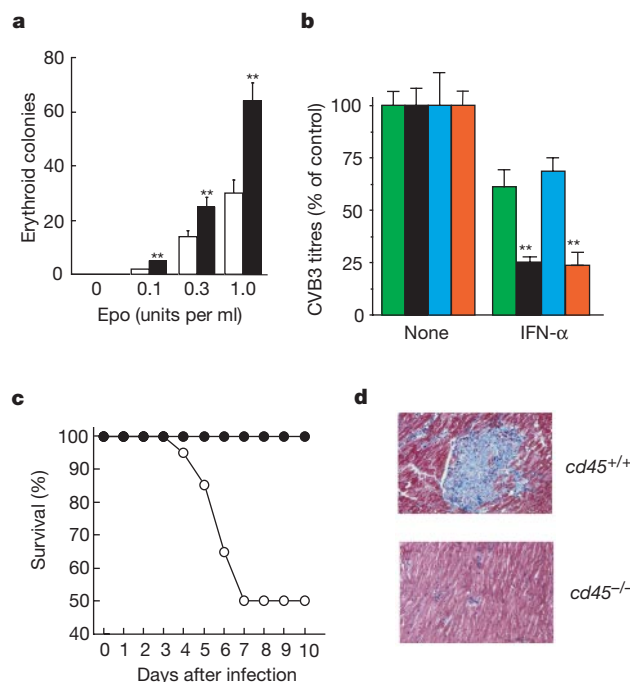


Figure 5 Enhanced erythroid colony formation and antiviral activity in the absence of CD45. **a**, *In vitro* EPO-dependent erythroid colony-forming ability of *cd45*^{+/-} (white columns) and *cd45*^{-/-} (black columns) bone-marrow progenitors. Mean numbers of erythroid colonies (±s.e.m.) were scored at day 7. Asterisks, *P* < 0.01 between *cd45*^{+/-} and *cd45*^{-/-} colonies. **b**, Parental Jurkat cells (green), *cd45*-deficient Jurkat cells (black), CD45-AS cells re-expressing wild-type CD45 (blue) and CD45-AS cells reconstituted with PTPase-dead CD45 (red) were left untreated (None) or treated with IFN-α before the

addition of CVB3. Asterisks, *P* < 0.01, wild-type or CD45-REC compared with CD45-AS or CD45-C828S cells. **c**, Survival of CVB3-infected *cd45*^{-/-} (filled circles) and *cd45*^{+/-} (open circles) littermate mice. CVB3-infected *cd45*^{+/-} mice showed signs of severe systemic illness, whereas *cd45*^{-/-} mice showed no signs of disease and had a 100% survival rate even at later time points (>60 days after infection). **d**, Histopathology of hearts of CVB3-infected *cd45*^{+/-} mice and *cd45*^{-/-} littermates (day 35 after initial infection). Masson staining. Original magnification × 200.

whether LAR is a JAK PTPase operating in non-haematopoietic cells and whether LAR-regulated JAK-STAT signalling has a role in glucose and/or fat metabolism.

The results of this study provide genetic, functional and biochemical evidence that the transmembrane protein tyrosine phosphatase CD45 can directly dephosphorylate and inactivate JAK-family kinases. CD45 negatively regulates JAK-STAT signalling in response to multiple cytokines, including IL-3, IL-4, EPO and IFN- α . Genetic inactivation of CD45 in cells enhanced cytokine-triggered proliferation, differentiation and antiviral activities. These data identify an unexpected role for the haematopoietic phosphatase CD45 as the elusive JAK phosphatase and a novel negative regulator of cytokine receptor signalling. □

Methods

Mice, cells and reagents

CD45-exon6 (ref. 2) and CD45-exon9 (ref. 3) mice have been previously generated and were maintained in accordance with institutional guidelines. BMMCs were cultured as described²⁶. Macrophages, thymocytes and splenic B cells were freshly isolated from *cd45* wild-type and *cd45*-mutant littermate mice²⁶. The CD45-negative human Jurkat cell lines CD45-AS (J-AS-1)¹¹ and J45.01 were cultured in RPMI medium plus 5% FCS (fetal calf serum). Phycoerythrin-, fluorescein isothiocyanate- or biotin-conjugated antibodies against pan-CD45, CD19, c-Kit (CD119), Gr1 and Mac1 (CD11b) (PharMingen) were used for flow cytometry. Antibodies against STAT3, STAT5, or PKB/Akt and STAT3 (Tyr 705), STAT3 (Ser 727), STAT5 (Tyr 694), STAT1 (Tyr 701), PKB/Akt (Ser 473 or Thr 308), ERK1/ERK2 (Thr 202 and Tyr 204), JAK1 (Tyr 1022 and Tyr 1023), Tyk2 (Tyr 1054 and Tyr 1055) and JAK2 (Tyr 1007 and Tyr 1008) were from QCB. The anti-phosphotyrosine (PY99) antibody and antisera against Lyn, JAK1, JAK2, Tyk2, JAK3 and c-Kit were from Santa Cruz Biotechnology. PP2, recombinant CD45, LAR, TC-PTP and lambda phosphatase were from Calbiochem. Antibodies against JAK3 and IL-3 receptor β -chain and purified JAK2 and Lyn were from Upstate Biotech. One unit of phosphatase activity was defined as the amount of enzyme that hydrolysed 1.0 nmol of *p*-nitrophenyl phosphate per minute at 30 °C and pH 7.0.

Immunoblotting, immunoprecipitation and transient expression of CD45

Murine BMMCs, splenic B cells, thymocytes, peritoneal macrophages and Jurkat T-cell lines were stimulated with various cytokines. For immunoprecipitations, lysates were incubated with antibodies conjugated with agarose or Sepharose beads for 2 h at 4 °C. Total cell lysates or immunoprecipitates were separated using SDS-polyacrylamide gel electrophoresis, transferred onto membranes and immunoblotted. COS cells were transfected using diethylaminoethyl-dextran transfection technique with wild-type or PTPase-dead CD45 (D1:C828s) as described¹¹.

Assays

For tyrosine phosphatase assays, immunoprecipitated Lyn, JAKs or STATs from lysates of 10^7 BMMCs (for Lyn, JAK1, JAK2, STAT3 and STAT5) or macrophages (for Tyk2) were washed in lysis buffer and phosphatase reaction buffer¹⁰. The amounts of Lyn and JAK2 in the immunoprecipitates used in Fig. 3a were 72 ng and 48 ng, respectively, as determined by silver staining. Recombinant CD45 was added to substrates in the presence or absence of 0.1 mM Na₃VO₄ followed by incubation at 30 °C for 20 min. Tyrosine phosphorylation was monitored by immunoblotting with phosphotyrosine-specific antibodies. The extent of tyrosine phosphorylation and amount of substrates were determined by quantifying the intensities of the immunoblots with NIH Image 1.61 software. Kinase activity in anti-Lyn or anti-JAK immunoprecipitates was assessed by an *in vitro* kinase assay with Raytide (Oncogen Science) as a substrate or using autophosphorylation²⁶. *In vitro* binding assays were performed as described¹⁰, using 1 μ g of soluble GST, GST-CD45-D2, or GST-PTPase-dead CD45 (D1:C828s). EMSAs were as described²⁷. The sequences of the EMSA probes were: STAT3, 5'-GATCCTTCTGGGAATTCCTAGATC-3'; STAT3 mutant, 5'-GATCCTTCTGGGCGCTCTAGATC-3'.

Proliferation, cell death, colony formation and CVB3 infections

For proliferation of BMMCs, cells were deprived of IL-3 for 12 h and then stimulated with IL-3 for 24 h and pulse-labelled with [³H]thymidine for 12 h. The percentage of cell death was determined by 7-amino-actinomycin D (7AAD) and staining with Annexin V. Haematopoietic precursors were isolated from the bone marrow of 6-week-old *cd45*^{-/-} and *cd45*^{+/-} mice and the formation of EPO-dependent erythroid colonies (BFU-E) and IL-3-dependent mixed neutrophil/macrophage myeloid colonies was determined. Total bone-marrow cells were plated in methylcellulose containing 5% fetal bovine serum. EPO (0.1, 0.3 and 1 unit ml⁻¹) or IL-3 (0.1 ng ml⁻¹) was added to the culture to induce erythroid or myeloid colony formation, respectively. The sizes and numbers of colonies were analysed 7 and 12 days after the start of culture.

Virus infections

For Coxsackievirus infections *in vitro*, Jurkat cells were pretreated with IFN- α (1,000 units ml⁻¹) for 4 h and inoculated in triplicate with 3×10^5 CVB3 for 1 h at 37 °C. Cells were then washed to remove excess IFN- α and excess CBV3 virus, and resuspended in RPMI medium containing 10% FBS. After 24 h, culture supernatants were analysed for the

presence of infectious CVB3 by using plaque-forming assays in HeLa cell monolayers. For CVB3 infections *in vivo*, the *cd45* mutation was backcrossed onto an A/J (H2^{klk}) background for five generations. A/J (H2^{klk}) mice are highly susceptible to acute and chronic CVB3 infections. Four-week-old mice were inoculated intraperitoneally with 10^5 plaque-forming units of CVB3. For lethality scores, *cd45*^{-/-} ($n = 25$) and *cd45*^{+/-} ($n = 22$) littermate mice were monitored daily after infection. CVB3-infected *cd45*^{+/-} mice showed signs of severe systemic illness, including lethargy, ruffled coats and anorexia, before death, but *cd45*^{-/-} mice failed to show these signs. At defined time points, mice were killed and organs were processed for histology. Sections from eight mice per group were stained with haematoxylin/eosin and Masson blue. In all experiments, the CVB3 strain Gauntt-Chow was used.

Received 5 September; accepted 17 November 2000.

1. Tonks, N. K. & Neel, B. G. From form to function: signaling by protein tyrosine phosphatases. *Cell* **87**, 365–368 (1996).
2. Kishihara, K. *et al.* Normal B lymphocyte development but impaired T cell maturation in CD45⁻ exon6 protein tyrosine phosphatase-deficient mice. *Cell* **74**, 143–156 (1993).
3. Byth, K. F. *et al.* CD45-null transgenic mice reveal a positive regulatory role for CD45 in early thymocyte development, in the selection of CD4⁺CD8⁺ thymocytes, and B cell maturation. *J. Exp. Med.* **183**, 1707–1718 (1996).
4. Trowbridge, I. S. & Thomas, M. L. CD45: an emerging role as a protein tyrosine phosphatase required for lymphocyte activation and development. *Annu. Rev. Immunol.* **12**, 85–116 (1994).
5. Thomas, M. L. The leukocyte common antigen family. *Annu. Rev. Immunol.* **7**, 339–369 (1989).
6. de Groot, R. P., Coffey, P. J. & Koenderman, L. Regulation of proliferation, differentiation and survival by the IL-3/IL-5/GM-CSF receptor family. *Cell Signal.* **10**, 619–628 (1998).
7. Leonard, W. J. & O'Shea, J. J. Jaks and STATs: biological implications. *Annu. Rev. Immunol.* **16**, 293–322 (1998).
8. Matsumura, I. *et al.* Transcriptional regulation of the cyclin D1 promoter by STAT5: its involvement in cytokine-dependent growth of hematopoietic cells. *EMBO J.* **18**, 1367–1377 (1999).
9. Hanke, J. H. *et al.* Discovery of a novel, potent, and Src family-selective tyrosine kinase inhibitor. Study of Lck- and FynT-dependent T cell activation. *J. Biol. Chem.* **271**, 695–701 (1996).
10. Felberg, J. & Johnson, P. Characterization of recombinant CD45 cytoplasmic domain proteins. Evidence for intramolecular and intermolecular interactions. *J. Biol. Chem.* **273**, 17839–17845 (1998).
11. McKeeney, D. W., Onodera, H., Gorman, L., Mimura, T. & Rothstein, D. M. Distinct isoforms of the CD45 protein-tyrosine phosphatase differentially regulate interleukin 2 secretion and activation signal pathways involving Vav in T cells. *J. Biol. Chem.* **270**, 24949–24954 (1995).
12. Ihle, J. N. Cytokine receptor signalling. *Nature* **377**, 591–594 (1995).
13. Petricoin, E. F. I. *et al.* Antiproliferative action of interferon- α requires components of T-cell-receptor signalling. *Nature* **390**, 629–632 (1997).
14. Parganas, E. *et al.* Jak2 is essential for signaling through a variety of cytokine receptors. *Cell* **93**, 385–395 (1998).
15. Marine, J. C. *et al.* SOCS3 is essential in the regulation of fetal liver erythropoiesis. *Cell* **98**, 617–627 (1999).
16. Rodig, S. J. *et al.* Disruption of the Jak1 gene demonstrates obligatory and nonredundant roles of the Jak1 in cytokine-induced biologic responses. *Cell* **93**, 373–383 (1998).
17. Kandolf, R., Canu, A. & Hofschneider, P. H. Coxsackie B3 virus can replicate in cultured human foetal heart cells and is inhibited by interferon. *J. Mol. Cell. Cardiol.* **17**, 167–181 (1985).
18. Ward, A. C., Tows, I. & Yoshimura, A. The Jak-Stat pathway in normal and perturbed hematopoiesis. *Blood* **95**, 19–29 (2000).
19. Yasukawa, H. *et al.* The JAK-binding protein JAB inhibits Janus tyrosine kinase activity through binding in the activation loop. *EMBO J.* **18**, 1309–1320 (1999).
20. Klingmüller, U., Lorenz, U., Cantley, L. C., Neel, B. G. & Lodish, H. F. Specific recruitment of SH-PTP1 to the erythropoietin receptor causes inactivation of JAK2 and termination of proliferative signals. *Cell* **80**, 729–738 (1995).
21. Ratei, R. *et al.* Immunophenotype and clinical characteristics of CD45-negative and CD45-positive childhood acute lymphoblastic leukemia. *Ann. Hematol.* **77**, 107–114 (1998).
22. Ozdemirli, M., Mankin, H. J., Aiesenberg, A. C. & Harris, N. L. Hodgkin's disease presenting as a solitary bone tumor. A report of four cases and review of the literature. *Cancer* **77**, 79–88 (1996).
23. Kung, C. *et al.* Mutations in the tyrosine phosphatase CD45 gene in a child with severe combined immunodeficiency disease. *Nature Medicine* **6**, 343–345 (2000).
24. Saad, M. J., Carvalho, C. R., Thirone, A. C. & Velloso, L. A. Insulin induces tyrosine phosphorylation of JAK2 in insulin-sensitive tissues of the intact rat. *J. Biol. Chem.* **271**, 22100–22104 (1996).
25. Ren, J. M. *et al.* Transgenic mice deficient in the LAR protein-tyrosine phosphatase exhibit profound defects in glucose homeostasis. *Diabetes* **47**, 493–497 (1998).
26. Liu, Q. *et al.* SHIP is a negative regulator of growth factor receptor-mediated PKB/Akt activation and myeloid cell survival. *Genes Dev.* **13**, 786–791 (1999).
27. Chaturvedi, P., Sharma, S. & Reddy, E. P. Abrogation of interleukin-3 dependence of myeloid cells by the *v-src* oncogene requires SH2 and SH3 domains which specify activation of STATs. *Mol. Cell. Biol.* **17**, 3295–3304 (1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Saunders for scientific editing; T. Mak for providing CD45 mutant mice; and T. Mak, M. Reth, W. Yeh, D. Barber, V. Stambolic, C. Mirtsos, K. Bachmaier, A. Oliveirados-Santos, M. Crackower, Y. Kong, N. Joza, I. Kozieradzki and Q. Liu for comments and advice. This work was supported by grants from the Canadian Institutes for Health Research (CIHR) and the National Cancer Institute (NCI) of Canada and Amgen to J.M.P., and from the NIH to D.M.R. J.M.P. holds a Canadian Research Chair in Cell Biology.

Correspondence and requests for materials should be addressed to T.S. (e-mail: tsasaki@rinshoken.or.jp) or J.M.P. (e-mail: jpenning@amgen.com).

MAD2 haplo-insufficiency causes premature anaphase and chromosome instability in mammalian cells

Loren S. Michel^{*†‡}, Vasco Liberal^{†‡}, Anupam Chatterjee[§], Regina Kirchwegger[‡], Boris Paschel^{||}, William Gerald[¶], Max Dobles[#], Peter K. Sorger[#], V. V. S. Murty[§] & Robert Benezra[‡]

^{*} Department of Medicine, and [¶] Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA
[‡] Cell Biology and Genetics Program, Memorial Sloan-Kettering Cancer Center and the Sloan-Kettering Division, Graduate School of Medical Sciences, Cornell University, New York, New York 10021, USA
[§] Department of Pathology, College of Physicians and Surgeons of Columbia University, New York, New York 10032, USA
^{||} Division of Hematology/Oncology and Robert H. Lurie Comprehensive Cancer Center, Northwestern University Medical School, Chicago, Illinois 60611, USA
[#] Biology Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[†] These authors contributed equally to this work

The mitotic checkpoint protein hMad2 is required to arrest cells in mitosis when chromosomes are unattached to the mitotic spindle¹. The presence of a single, lagging chromosome is sufficient to activate the checkpoint, producing a delay at the metaphase–anaphase transition until the last spindle attachment is made². Complete loss of the mitotic checkpoint results in embryonic lethality owing to chromosome mis-segregation in various organisms^{3–6}. Whether partial loss of checkpoint control leads to more subtle rates of chromosome instability compatible with cell viability remains unknown. Here we report that deletion of one *MAD2* allele results in a defective mitotic checkpoint in both human cancer cells and murine primary embryonic fibroblasts. Checkpoint-defective cells show premature sister-chromatid separation in the presence of spindle inhibitors and an elevated rate of chromosome mis-segregation events in the absence of these agents. Furthermore, *Mad2*^{+/-} mice develop lung tumours at high rates after long latencies, implicating defects in the mitotic checkpoint in tumorigenesis.

Genetic instability in cancers occurs frequently with whole chromosomes, referred to as chromosome instability (CIN)⁷. CIN cells become aneuploid (that is, they acquire abnormal chromosome numbers), a hallmark of cancer that is associated with aggressive tumour behaviour and a poor prognosis⁸. Aneuploidy may facilitate tumorigenesis or tumour progression through the loss of tumour-suppressor gene function. The mechanisms responsible for the development of aneuploidy in cancers remain unclear. Defects in the mitotic checkpoint have been correlated with CIN in human cancer cells but have not yet been shown to be a direct cause of the CIN phenotype⁹. We therefore investigated whether the loss of *MAD2* through targeted deletion in a cancer cell line that is chromosomally stable would result in its transformation into a CIN line. We chose the Hct-116 human colon carcinoma cell line because it has a normal mitotic checkpoint and is amenable to gene targeting^{9–11}.

We generated two independent Hct-*MAD2*^{+/-} clones for analysis (For details of the vector design and genomic analysis, see Supplementary Information). We exposed asynchronous Hct-*MAD2*^{+/-} cells to nocodazole and then measured the mitotic index by staining them with MPM-2, a mitosis-specific antigen. Sixty per cent of wild-type cells arrest by 18 h, whereas only 10% of mutant cells do so (Fig. 1a). Consistent with the loss of the mitotic checkpoint, 77% of

mutant cells proceed into a subsequent S phase, as assayed by DNA content, compared with 36% of wild-type cells (Fig. 1b). The checkpoint defect of *MAD2*^{+/-} cells was confirmed by assaying cyclin-B-associated Cdc2 kinase activity of G1/S synchronized cell populations that had been released into nocodazole. Wild-type cells show high levels of Cdc2 kinase activity for 24 h, whereas the kinase activity of mutant cells falls rapidly after 12 h in nocodazole (Fig. 1c, d). This fall in kinase activity is accompanied by the acquisition of 8N DNA content (where N is the amount of DNA in a haploid cell), as mutant cells proceed through the next cell cycle (Fig. 1e). After 24 h, most of the wild-type cells have rounded up and detached from the culture dish, whereas mutant cells remain adherent as they continue to cycle (Fig. 1f, g). On prolonged exposure to nocodazole (over 48 h), mutant cells become multinucleated with a polyploid DNA content (Fig. 1h).

To ensure that the phenotype observed in Hct-*MAD2*^{+/-} cells was not the result of the transfection or selection process, two selected clones that had not recombined homologously were challenged with nocodazole and found by MPM-2 staining to arrest in mitosis identically to Hct-116 non-transfected controls (data not shown). Immunoblots against *MAD2* with polyclonal anti-sera show one species with a relative molecular mass of 27,000 (*M_r* 27K) in both Hct-*MAD2*^{+/-} clones, indicating that there is not a dominant-negative effect of a partial protein product. Quantitative iodinated immunoblots normalized to Cdk2 show that Hct-*MAD2*^{+/-} cells express 70% of the amount of *MAD2* protein expressed in wild-type cells, suggesting partial compensation of protein levels in the absence of one allele (see Fig. 2, Supplementary Information). We found that the expressed allele is wild type by polymerase chain reaction with reverse transcription of RNA and subsequent sequence analysis. Finally, the checkpoint status of ten independent subclones derived from the Hct-116 parental line was analysed and determined to be normal, which excludes clonal variation as the cause of the checkpoint defect in Hct-*MAD2*^{+/-} clones (data not shown).

MAD2 arrests cells in metaphase by associating with the anaphase-promoting complex (APC), thereby inhibiting its ubiquitin ligase activity^{12–14}. On release from the checkpoint, progression into anaphase and sister-chromatid separation requires APC^{Cdc20}-dependent degradation of Pds1, or securin^{15,16}. Checkpoint-challenged mutant cells containing catalytically active APC are expected to degrade securin prematurely; indeed, after release from a double thymidine block into nocodazole, Hct-*MAD2*^{+/-} cells fail to accumulate the high levels of securin seen in wild-type cells (Fig. 2a). As the mutant cells exit mitosis, the observed loss of Cdc2 activity is not accompanied by a concomitant decrease in cyclin B protein levels (data not shown). As phosphorylation of CDC28 in yeast is a potential mechanism of mitotic exit in checkpoint-challenged cells¹⁷ (see also ref. 18), we performed immunoblots using antisera specific for phosphorylated forms of Tyr 15 on Cdc2. We observed no increases in Tyr 15 phosphorylation, nor were we able to detect by co-immunoprecipitation any significant decreases in cyclin-B-associated Cdc2 protein levels between 12 and 16 h when the decline in kinase activity is most pronounced (Fig. 2b and data not shown). Further studies are required to determine the precise mechanism of mitotic exit in checkpoint-challenged mammalian cells.

Sister chromatids that have separated prematurely are a hallmark of a defective mitotic checkpoint and have been observed in yeast *mad2* mutants and *Drosophila bub1* mutants^{3,16,19}. Therefore, we looked for evidence of precocious anaphase in checkpoint-deficient cells by preparation of metaphase spreads in the presence of colcemid, a spindle inhibitor. Roughly 20% of mitotic figures from the Hct-*MAD2*^{+/-} cells showed prematurely separated sister chromatids, compared with 1% in wild-type cells (Fig. 3a–d).

We performed chromosome counts on metaphase spreads to test the importance of the mitotic checkpoint on chromosome stability. Hct-*MAD2*^{+/-} cells showed an 80% increase in the frequency of

aneuploid metaphases relative to wild-type Hct-116 cells, which have a relatively stable (45,X) karyotype (Fig. 3e). We did not perform karyotype analysis on cells containing 45 chromosomes, which possibly resulted in an underestimation of the total number of mis-segregation events. Using interphase fluorescence *in situ* hybridization (FISH) with chromosome-specific anti-centromeric probes for several chromosomes statistically significant differences in chromosome numbers ($P < 0.008$ to $P < 0.00001$, by Fisher exact test) were identified between wild-type and Hct-*MAD2*^{+/-} cells (Fig. 4a, b). We did not identify differences for chromosome 17 (data not shown). The mechanism of chromosome mis-segregation in checkpoint-defective Hct-*MAD2*^{+/-} cells may be due to premature anaphase even in the absence of spindle inhibitors, but other molecular explanations are possible.

We measured the rates of chromosome loss on clonal populations

to determine whether Hct-*MAD2*^{+/-} cells possess a perpetual CIN phenotype, or whether they have acquired an abnormal karyotype as a result of rare events propagated in subsequent generations. We performed chromosome counts on two independently derived wild-type and four Hct-*MAD2*^{+/-} clones after 25 generations. We observed a roughly 100% increase in the rate of chromosome losses in Hct-*MAD2*^{+/-} cells (Fig. 4c). Notably, one of the mutant clones, which has high rates of chromosome instability, has a modal karyotype of 47 chromosomes, possibly reflecting the high degree of karyotypic heterogeneity in the parental Hct-*MAD2*^{+/-} population.

To determine whether the haplo-insufficient effect seen in transformed cells also occurs in primary cells, we studied metaphase spreads from murine embryonic fibroblasts (MEFs) derived from the *Mad2*^{+/-} mouse⁵. MEFs from four out of five *Mad2*^{+/-} mice showed high frequencies of premature sister-chromatid separation

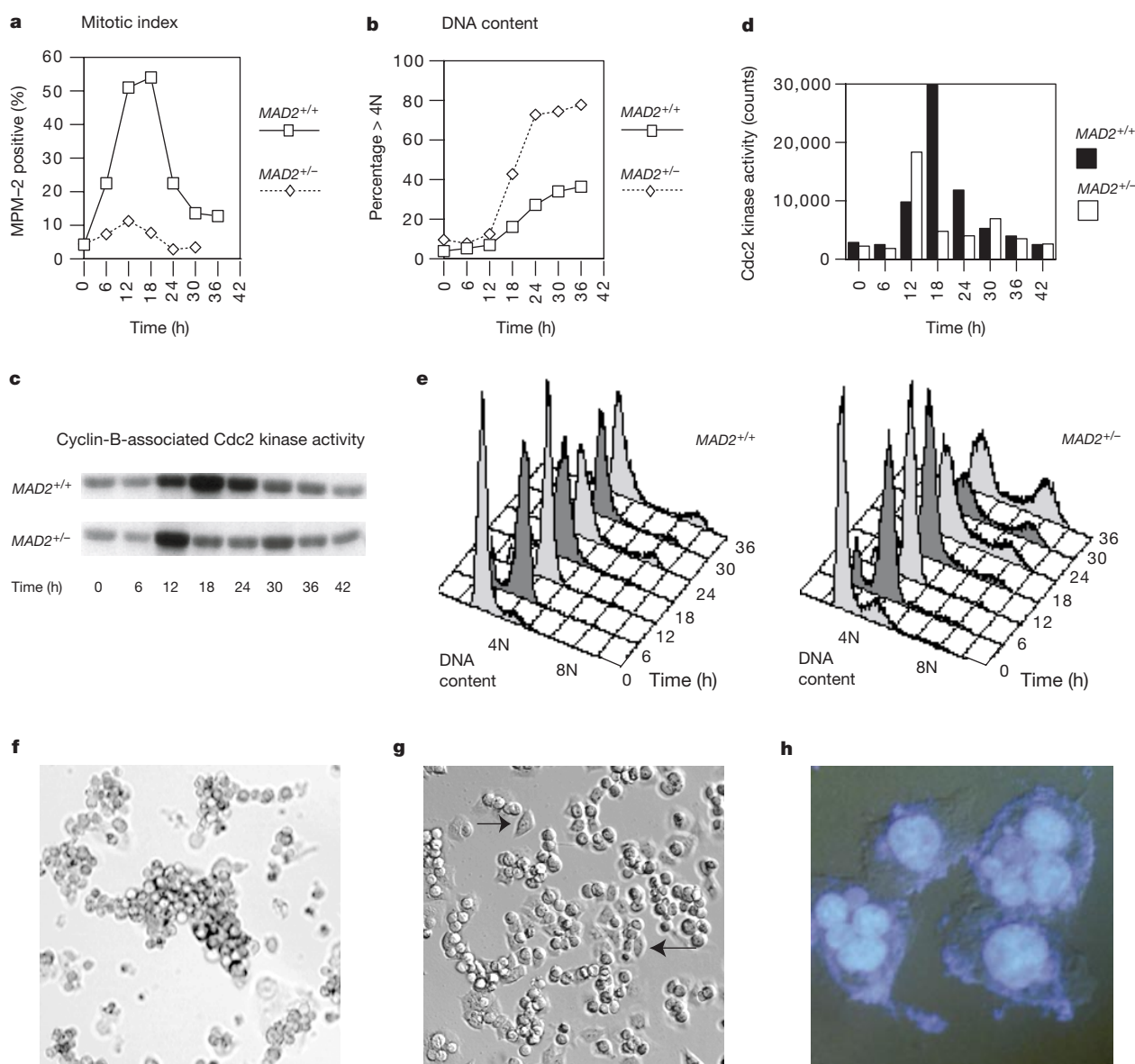


Figure 1 Hct-*MAD2*^{+/-} cells have a mitotic checkpoint defect. **a**, **b**, FACS profile of unsynchronized human Hct-116 and Hct-*MAD2*^{+/-} cells treated with nocodazole and stained with MPM-2 (**a**) and with propidium iodide (**b**). **c**, Cyclin-B-associated Cdc2 kinase activity on histone H1 substrate of Hct-116 versus Hct-*MAD2*^{+/-} cells synchronized in G1/S by double thymidine block and then released into nocodazole. **d**, Cyclin-B-associated Cdc2 kinase activity from the same time course. **e**, FACS profile of propidium iodide-stained cells from the same time course. The 42-h time point is not shown because

of its similarity to the 36-h time point. **f**, DIC microscopy images of unsynchronized Hct-116 cells exposed to nocodazole for 24 h showing an almost complete absence of adherent cells. **g**, DIC microscopy images of Hct-*MAD2*^{+/-} cells that remain attached to the tissue-culture dish after treatment for 24 h. Arrows show adherent cells. **h**, Hct-*MAD2*^{+/-} cells exposed to nocodazole for 52 h and analysed by DAPI staining and immunofluorescence microscopy become multinucleated. Magnification of **f**–**h** $\times 1,000$.

relative to wild-type control cultures (Table 1; see also Supplementary Information). We performed chromosome counts on three of the four MEF *Mad2*^{+/-} cultures and in all cases there was a significant increase in the number of aneuploid cells relative to the controls (Table 1). The percentage of aneuploid cells correlates with the severity of premature sister-chromatid separation, suggesting a direct relationship between the severity of loss of checkpoint control and chromosome mis-segregation (Table 1). We note that *Mad2*^{+/-} mice develop normally, and blastocysts from *Mad2* heterozygote animals display no evidence of an impaired checkpoint as assayed by reactivity to phospho-histone 3 antibody, but premature sister-chromatid separation was not examined in these experiments⁵.

Importantly, a high frequency of papillary lung adenocarcinomas, an extremely rare tumour in most wild-type mouse strains, is identifiable in *Mad2*^{+/-} mice killed between 18 and 19 months of age, which suggests a role for defects in the mitotic checkpoint in

tumorigenesis (Table 2)²⁰. The background rate of lymphomas seen in ageing mice is unchanged in the heterozygote mice. Why the lung is particularly affected by loss of mitotic checkpoint control, the cause of the latency and the rate of chromosome mis-segregation in these tumours are currently being investigated, and the results are likely to clarify the role of mitotic checkpoints in tumour initiation and progression.

The molecular basis of the CIN phenotype has, until recently, remained obscure^{21,22}. Our results indicate that partial loss of the mitotic checkpoint may not only be a direct cause of chromosome instability, but may also contribute to tumorigenesis. Defects in the mitotic checkpoint are common in human cancer and have been identified in colon, lung and breast cancer cell lines¹. Although *MAD2* mutations have not been detected in colon and lung cancer cell lines with mitotic checkpoint defects^{23,24}, we show here that subtle differences in *MAD2* protein levels markedly alter checkpoint function. Therefore, inactivation of one *MAD2* allele, either by

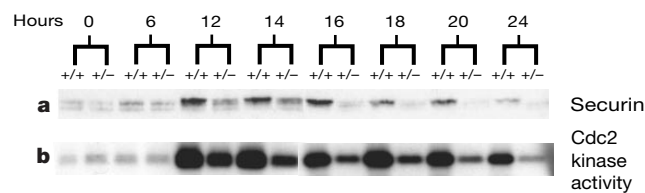


Figure 2 Inappropriate degradation of securin in Hct-*MAD2*^{+/-} cells. **a**, Immunoblot analysis of total cellular protein extract obtained over 24 h from Hct-116 and Hct-*MAD2*^{+/-} cells synchronized by double thymidine block in G1/S and released into nocodazole.

b, Cyclin-B-associated Cdc2 kinase activity on histone H1 from the same time course. Synchronization and progression through the cell cycle was confirmed by FACS of propidium-iodide-stained cells (data not shown).

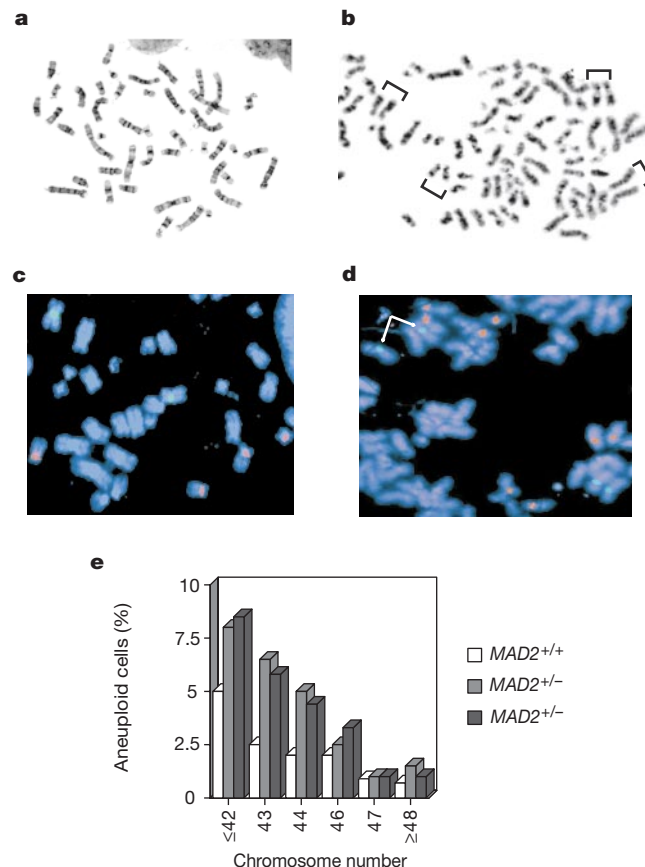


Figure 3 Karyotype analysis of genomic instability in Hct-*MAD2*^{+/-} cell lines. **a**, Normal G-banded metaphase spread from Hct-116 cell lines. **b**, Premature sister-chromatid separation from one Hct-*MAD2*^{+/-} cell line. Brackets show sister chromatids lying separated and adjacent to one another. All of the sister-chromatid pairs are prematurely separated in mitotic figures that show this phenotype. FISH analysis using chromosome-

specific FISH probes for CEP 12 (red), CEP 16 (red) and CEP 7 (green), on mitotic figures from the Hct-116 cell line (**c**) and Hct-*MAD2*^{+/-} cell line (**d**). Magnification of **c** and **d**, $\times 1,000$. Arrows indicate a prematurely separated sister-chromatid pair with a third chromatid lying between the two. **e**, Distribution of chromosome numbers of Hct-116 and two Hct-*MAD2*^{+/-} cell lines. At least 400 metaphases were counted for each cell line.

Table 1 Chromosome analysis of mouse embryonic fibroblasts

	Percentage of metaphases with premature sister-chromatid separation <i>n</i> ≥ 100	Percentage of aneuploid cells
<i>Mad2</i> ^{+/+}	7	16 (<i>n</i> = 102)
<i>Mad2</i> ^{+/-}	30	40 (<i>n</i> = 42)
<i>Mad2</i> ^{-/-}	36	57 (<i>n</i> = 60)
<i>Mad2</i> ^{+/-}	57	75 (<i>n</i> = 33)

epigenetic silencing or deletion, would be sufficient to create a haplo-insufficient effect and the loss of mitotic checkpoint control. Of note, loss of heterozygosity around 4q27, to which *MAD2* has been mapped, has been identified in several different tumour types^{25–28}.

Our results also provide a new cytogenetic marker (premature sister-chromatid separation) that should guide screening efforts for the identification of mitotic checkpoint defects in primary tumour samples. Preliminary analysis in primary human cancers has identified such defects (A.C. and V.V.V.S.M., unpublished observations), but further investigation is needed as to whether these are due to subtle decreases in *MAD2* levels, or other components of the spindle assembly checkpoint. The precise involvement that a partial loss of mitotic checkpoint function has in the development, progression and management of cancer remains to be determined. As *MAD2*^{+/-} cells become highly aneuploid and continue to cycle after prolonged exposure to spindle inhibitors, checkpoint-defective tumours may be more aggressive and resistant to anti-mitotic chemotherapeutics. Furthermore—and more fundamentally—the long latency period

Table 2 Tumour frequency in ageing mice

	Lymphomas	Lung tumours
<i>Mad2</i> ^{+/+}	7.4% (2/27)	0% (0/27)
<i>Mad2</i> ^{+/-}	9.8% (5/51)	27.5% (14/51)

Wild-type and *Mad2*^{+/-} mice were killed at the age of 18–19 months.

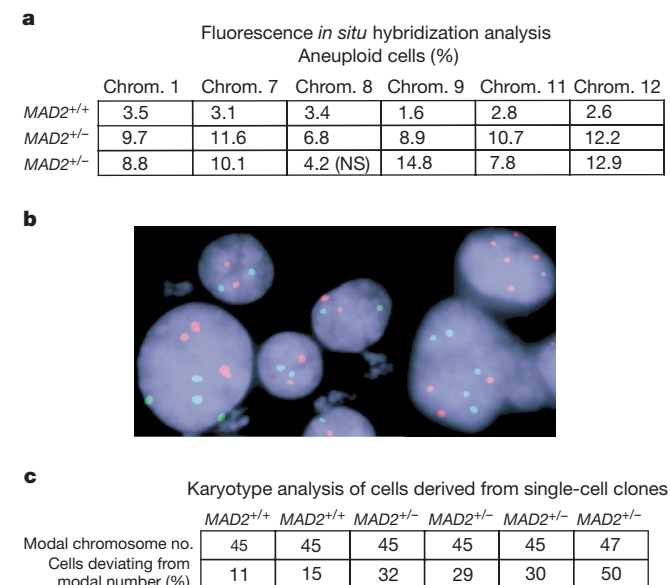


Figure 4 Interphase FISH analysis using chromosome-specific FISH probes on Hct-116 and Hct-*MAD2*^{+/-} cells. **a**, Summary of six different experiments. 400–600 cells were scored for aneuploidy for each cell line. NS; not significant. **b**, Microscopy image of interphase FISH on Hct-*MAD2*^{+/-} cells using probes for chromosomes 7 (green) and 8 (red) showing three normal and three aneuploid cells. Magnification × 1,000. **c**, Chromosome counts from metaphase spreads on cell populations derived from single cell Hct-116 and Hct-*MAD2*^{+/-} clones after 25 generations. Roughly 100 metaphases were counted for each cell line.

essential for tumour formation in the *Mad2*^{+/-} mice suggests that cooperative mutations at additional oncogenic or tumour-suppressor loci are required to drive transformation in checkpoint-defective cells. Further analysis of the *Mad2*^{+/-} mice should identify these putative genetic interactions. □

Methods

Cell-cycle analysis and Cdc2 kinase assays

We placed asynchronous cells in 200 ng ml⁻¹ nocodazole, and then fixed in 70% ethanol at the given time points. Cells were stained with anti-MPM-2 mouse monoclonal antibody (Upstate Biotechnology), washed and stained with goat anti-mouse fluorescein isothiocyanate (Sigma) and propidium iodide. We performed cell synchronization and H1 kinase assays as described¹⁴. Detached cell populations were collected and combined with adherent cells for all experiments.

Protein and immunoblot analysis

We prepared total cellular protein extracts as described¹⁴. For securin, cyclin B and anti-Tyr 15 Cdc2 immunoblots, 30 µg of total cellular extract was resolved by SDS-polyacrylamide gel electrophoresis, transferred to nitrocellulose, and probed with rabbit anti-securin polyclonal antibodies, rabbit anti-cyclin B polyclonal antisera, (Santa Cruz), or phospho-specific anti-Cdc2 (Tyr 15) rabbit polyclonal antibodies (New England Biolabs). Horseradish-peroxidase were conjugated donkey anti-rabbit antisera (Amersham) used as secondary antibody, and immunodetection was accomplished by enhanced chemiluminescence (Amersham).

Methods for karyotype, FISH and microscopy

We performed metaphase spreads and Giemsa staining as described²⁹. We carried out FISH using chromosome-specific centromeric probes obtained from Vysis, in accordance with the manufacturer's protocol. We analysed slides stained with DAPI counterstain with a Nikon Eclipse 600 microscope attached to an applied imaging cytovision imaging system. We obtained mouse embryonic fibroblasts from 13.5-day-old embryos and used them at passage 3 for metaphase spread preparation as above. For DIC and immunofluorescence microscopy, live cells or cells fixed with 2% paraformaldehyde and stained with 1 µg ml⁻¹ DAPI were visualized with an Axiovert 100M microscope (Zeiss) using DIC imaging.

Animal breeding and tumour analysis

Mad2^{+/-} mice were screened and identified as described³. Tissues were collected, fixed in formalin and embedded in paraffin blocks before staining with haematoxylin and eosin, and microscopic analysis.

Received 4 July; accepted 23 October 2000.

- Li, Y. & Benezra, R. Identification of a human mitotic checkpoint gene: *hSMAD2*. *Science* **274**, 246–248 (1996).
- Rieder, C. L., Cole, R. W., Khodjakov, A. & Sluder, G. The checkpoint delaying anaphase in response to chromosome monoorientation is mediated by an inhibitory signal produced by unattached kinetochores. *J. Cell Biol.* **130**, 941–948 (1995).
- Basu, J. *et al.* Mutations in the essential spindle checkpoint gene *bub1* cause chromosome missegregation and fail to block apoptosis in *Drosophila*. *J. Cell Biol.* **146**, 13–28 (1999).
- Kitagawa, R. & Rose, A. M. Components of the spindle-assembly checkpoint are essential in *Caenorhabditis elegans*. *Nature Cell Biol.* **1**, 514–521 (1999).
- Dobles, M., Liberal, V., Scott, M., Benezra, R. & Sorger, P. Chromosome missegregation and apoptosis in mice lacking the mitotic checkpoint protein *Mad2*. *Cell* **101**, 635–645 (2000).
- Kalitsis, P., Earle, E., Fowler, K. J. & Choo, K. H. *Bub3* gene disruption in mice reveals essential mitotic spindle checkpoint function during early embryogenesis. *Genes Dev.* **14**, 2277–2282 (2000).
- Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
- DeVita, V., Hellman, S. & Rosenberg, S. (eds) *Cancer—Principles and Practice* (Lipincott-Raven, New York, 1997).
- Cahill, D. P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392**, 300–303 (1998).
- Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386**, 623–627 (1997).
- Waldman, T., Lengauer, C., Kinzler, K. W. & Vogelstein, B. Uncoupling of S phase and mitosis induced by anticancer agents in cells lacking p21. *Nature* **381**, 713–716 (1996).
- Li, Y., Gorbea, C., Mahaffey, D., Rechsteiner, M. & Benezra, R. *MAD2* associates with the cyclosome/anaphase-promoting complex and inhibits its activity. *Proc. Natl Acad. Sci. USA* **94**, 12431–12436 (1997).
- Fang, G., Yu, H. & Kirschner, M. W. The checkpoint protein *MAD2* and the mitotic regulator *CDC20* form a ternary complex with the anaphase-promoting complex to control anaphase initiation. *Genes Dev.* **12**, 1871–1883 (1998).
- Wassmann, K. & Benezra, R. *Mad2* transiently associates with an APC/p55Cdc complex during mitosis. *Proc. Natl Acad. Sci. USA* **95**, 11193–11198 (1998).
- Zou, H., McGarry, T. J., Bernal, T. & Kirschner, M. W. Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. *Science* **285**, 418–422 (1999).
- Alexandru, G., Zachariae, W., Schleiffer, A. & Nasmyth, K. Sister chromatid separation and chromosome re-duplication are regulated by different mechanisms in response to spindle damage. *EMBO J.* **18**, 2707–2721 (1999).
- Rudner, A. D. & Murray, A. W. The spindle assembly checkpoint. *Curr. Opin. Cell Biol.* **8**, 773–780 (1996).

18. Rudner, A. D., Hardwick, K. G. & Murray, A. W. Cdc28 activates exit from mitosis in budding yeast. *J. Cell Biol.* **149**, 1361–1376 (2000).
19. Minshull, J. et al. Protein phosphatase 2A regulates MPF activity and sister chromatid cohesion in budding yeast. *Curr. Biol.* **6**, 1609–1620 (1996).
20. Tuveson, D. A. & Jacks, T. Modeling human lung cancer in mice: similarities and shortcomings. *Oncogene* **18**, 5318–5324 (1999).
21. Zhou, H. et al. Tumour amplified kinase STK15/BTAK induces centrosome amplification, aneuploidy and transformation. *Nature Genet.* **20**, 189–193 (1998).
22. Spruck, C. H., Won, K. A. & Reed, S. I. Deregulated cyclin E induces chromosome instability. *Nature* **401**, 297–300 (1999).
23. Takahashi, T. et al. Identification of frequent impairment of the mitotic checkpoint and molecular analysis of the mitotic checkpoint genes, *hMAD2* and *p55CDC*, in human lung cancers. *Oncogene* **18**, 4295–4300 (1999).
24. Cahill, D. P. et al. Characterization of *MAD2B* and other mitotic spindle checkpoint genes. *Genomics* **58**, 181–187 (1999).
25. Krishnan, R. et al. Map location and gene structure of the *Homo sapiens* mitotic arrest deficient 2 (*MAD2L1*) gene at 4q27. *Genomics* **49**, 475–478 (1998).
26. Rashid, A. et al. Genetic alterations in hepatocellular carcinomas: association between loss of chromosome 4q and p53 gene mutations. *Br. J. Cancer* **80**, 59–66 (1999).
27. Dohner, H., Bloomfield, C. D., Frizzera, G., Freestadt, J. & Arthur, D. C. Recurring chromosome abnormalities in Hodgkin's disease. *Genes Chromosomes Cancer* **5**, 392–398 (1992).
28. Shivapurkar, N. et al. Multiple regions of chromosome 4 demonstrating allelic losses in breast carcinomas. *Cancer Res.* **59**, 3576–3580 (1999).
29. Verma, R. & Babu, A. *Human Chromosomes: Principles and Techniques* (McGraw Hill, New York, 1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Kirschner for the anti-securin antibody. A.C. performed all cytogenetic and FISH experiments. This work was supported by an NRSA-NIH grant (L.M.); PRAXISXXI, Portugal (V.L.); International Union Against Cancer (ACS-IFB) Fellowship (A.C.); and grants from the NIH (R.B., B.P., P.K.S. and V.V.S.M.). We also thank A. Hata and G. Farmer for technical advice, and H. Thaler for statistical analysis.

Correspondence and requests for materials should be addressed to R.B. (e-mail: r-benezra@ski.mskcc.org).

Pre-meiotic S phase is linked to reductional chromosome segregation and recombination

Yoshinori Watanabe^{*†‡}, Shihori Yokobayashi^{*}, Masayuki Yamamoto^{*} & Paul Nurse[‡]

^{*} Department of Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Hongo, Tokyo 113-0033, Japan

[†] PRESTO, Japan Science and Technology Corporation, Kawaguchi, Saitama 332-0012, Japan

[‡] Imperial Cancer Research Fund, 44 Lincoln's Inn Field, London WC2A 3PX, UK

Meiosis is initiated from G1 of the cell cycle and is characterized by a pre-meiotic S phase followed by two successive nuclear divisions. The first of these, meiosis I, differs from mitosis in having a reductional pattern of chromosome segregation^{1,2}. Here we show that meiosis can be initiated from G2 in fission yeast cells by ectopically activating the meiosis-inducing network. The subsequent meiosis I occurs without a pre-meiotic S phase and with decreased recombination, and exhibits a mitotic pattern of equational chromosome segregation. The subsequent meiosis II results in random chromosome segregation. This behaviour is similar to that observed in cells lacking the meiotic cohesin Rec8 (refs 3, 4), which becomes associated with chromosomes at G1/S phase, including the inner centromere, a region that is probably critical for sister-centromere orientation⁵. If the expression of Rec8 is delayed to S phase/G2, then the centromeres behave equationally. We propose that the presence of Rec8 in chromatin is required at the pre-meiotic S phase to construct centromeres that behave

reductionally and chromosome arms capable of a high level of recombination, and that this explains why meiosis is initiated from G1 of the cell cycle.

The meiosis-inducing regulatory network is well characterized in fission yeast^{6–9}. We have found that this network is activated only at G1, which explains why meiosis is initiated only during this phase of the cell cycle (see Supplementary Information). On the basis of these findings, we have developed a method of inducing synchronous meiosis from G2 (G2-exit meiosis) by ectopically activating the meiosis-inducing regulatory network during G2 (see Methods), and have compared G2-exit meiosis with meiosis induced from G1 (G1-exit meiosis) (Fig. 1). The cells entering meiosis from G1 underwent a pre-meiotic S phase, meiotic nuclear divisions and sporulation as expected. By contrast, the cells entering meiosis from G2, after mitotic S phase, failed to undergo DNA replication, but did undergo meiotic nuclear divisions and sporulation (Fig. 1a, b, d).

To characterize G2-exit meiosis, we first monitored spindle pole bodies and showed that they remained unseparated until 4 h, indicating that, as in G1-exit meiosis, the spindle was not formed until just before meiosis I (Fig. 1b, and data not shown). Second, we induced meiosis from G2 in *mei4Δ* mutant. The *mei4* gene is required specifically for meiosis I but is not required for mitosis¹⁰. The *mei4Δ* cells undergoing G2-exit meiosis all arrested with a single nucleus, establishing that the nuclear divisions were meiotic in character rather than mitotic (data not shown). Third, we examined transcripts of the meiosis-specific genes *rec12*, *rec6*, *rec8*, *spo5*, *mei4* and *spo6*, and found that all were induced during both G1-exit and G2-exit meioses (Fig. 1c). Two genes required for DNA replication, *cdc18* and *cdc22*, were induced only during G1-exit meiosis (Fig. 1c), which was consistent with the fact that G2-exit meiotic cells do not undergo a pre-meiotic S phase. Thus, the normal meiotic transcriptional programme is activated during G2-exit meiosis. Last, we assessed intragenic recombination levels by using two closely linked *ade6* mutations¹¹. Recombination levels during G1-exit meiosis were similar to those observed during a normal physiological meiosis, but were 20-fold higher than those observed during G2-exit meiosis. The latter level was similar to, although somewhat higher than, the recombination level observed during a mitotic cell cycle. Measurement of intergenic recombination between *lys1* and *leu2* yielded similar results (see Supplementary Information). We conclude that G2-exit meiosis shows several characteristics of a normal meiosis, although recombination levels are lower.

We next investigated chromosome segregation. The *cdc2^{ts}* mutation blocks meiosis II, leading to the generation of dyads containing two spores¹², but the *cdc2^{ts}* allele does not impair meiotic recombination or reductional division at meiosis I (ref. 4, and data not shown). The cessation of meiosis before the second division allows us to distinguish between reductional and equational chromosome segregation during meiosis I (ref. 4). Spores from dyads showed over 80% viability in both G1-exit and G2-exit meioses, but spores from tetrads were only 14% viable in G2-exit meiosis in comparison with 86% viable in G1-exit meiosis (Fig. 1e). Staining of nuclei in G2-exit meiosis with 4,6-diamidino-2-phenylindole (DAPI) indicated that chromosome segregation was equal during meiosis I but unequal during meiosis II (Fig. 1d). To assess whether meiosis I was reductional or equational, the segregation of two centromere-linked markers, *lys1* and *ade6*, was monitored in dyads. In G1-exit meiosis, segregation was predominantly +/+, –/–, which was indicative of reductional segregation (Fig. 1f). The occurrence of some +/- dyads can be explained by recombination between the markers and the centromeres; recombination in the dyads would be expected to be 8% with *lys1* marker (4 cM from the centromere of chromosome I) and 32% with *ade6* marker (16 cM from the centromere of chromosome III). In contrast, segregation during G2-exit meiosis was entirely equational, as indicated by the presence of exclusively

+/-, +/- dyads (Fig. 1f). The lack of +/+, -/- dyads reflects the lower rate of recombination seen in G2-exit meiosis. We conclude that G2-exit meiosis results in equational instead of reductional segregation during meiosis I. Equational segregation at meiosis I might result in a loss of cohesion between chromatids. This was confirmed experimentally by monitoring the green fluorescent protein (GFP) fluorescence associated with the centromere-linked *lys1* locus (Cen1-GFP)¹³ (Fig. 1g). Lack of cohesion then leads to random chromosome segregation during meiosis II (Fig. 1h).

As observed with the *spo11* mutant in budding yeast¹⁴, the depletion of meiotic recombination by the mutation of *rec12* (the *spo11* homologue in fission yeast) does not lead to the separation of sister chromatids during meiosis I (Y.W., unpublished observations). Therefore the equational division observed in G2-exit meiosis is not the indirect effect of a decrease in recombination. We have found previously that deletion of the meiotic cohesin Rec8

leads to a defective meiosis in which the first nuclear division is equational, and meiotic recombination levels at the *ade6* locus or between *lys1* and *leu2* loci are approximately 20-fold lower than wild type (ref. 4, and data not shown). These phenotypes are identical to those found here in G2-exit meiosis. However, a substantial amount of Rec8 was produced in G2-exit meiosis and became localized in the chromatin at centromeres and arm regions during meiotic prophase (Fig. 1i). It therefore seems that Rec8 is required during the pre-meiotic S phase to promote recombination and reductional chromosome segregation. If this is so, then Rec8 must be incorporated into chromatin before the pre-meiotic S phase in normal (G1-exit) meiosis. To explore this possibility, we examined the association of Rec8 with chromatin by chromatin immunoprecipitation (CHIP) assays during synchronous G1-exit meiosis (Fig. 2). We immunoprecipitated Rec8-GFP and DNA complexes with anti-GFP antibodies and subsequently performed quantitative

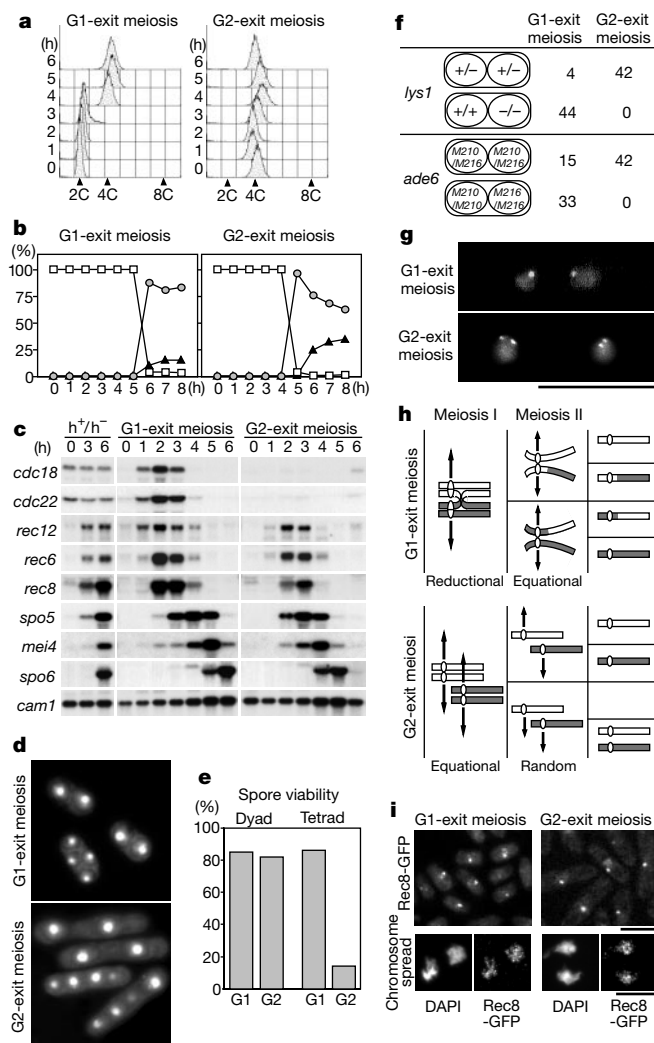


Figure 1 Comparison between G1-exit and G2-exit meiosis. Synchronous G1-exit and G2-exit meioses were induced by using a *pat1^{ts} cdc2^{ts} h⁺/h⁺* diploid strain. **a**, Flow cytometric determination of the cellular DNA content. **b**, Number of nuclei per cell counted after staining with DAPI: squares, one nucleus; circles, two nuclei; triangles, three or four nuclei. Note that only some cells underwent meiosis II because of the *cdc2^{ts}* mutation. **c**, RNA analysed by northern blotting. Meiotic *h⁺/h⁺* wild-type cells were used as a reference. The genes analysed were as follows: *cdc18* and *cdc22*, which are essential for DNA replication^{23,24}; *rec12*, *rec6* and *rec8*, which are required for meiotic recombination²⁵; *mei4*, which is essential for meiosis I; *spo5* and *spo6*, which are essential for spore formation¹⁰; and *cam1*, encoding calmodulin, used as a loading control. **d**, The cells that underwent G1-exit or G2-exit meiosis were stained with DAPI.

Note that in G2-exit meiosis the DNA distribution between tetrad spores is unequal, whereas that between dyad spores is equal. **e**, The tetrad and dyad spores were dissected and examined for viability. **f**, Full viable dyads were examined for segregation of the centromere-linked markers. **g**, Cells in G2-exit meiosis lost chromatid cohesion after meiosis I, as shown by separated Cen1-GFP in each nucleus. Cells in G1-exit meiosis showed a co-localized signal. Cells were arrested before meiosis II by introducing the *mes1* mutation. **h**, Schematic drawing of the behaviour of homologous chromosomes (white and black) in G1-exit and G2-exit meioses. **i**, Localization of Rec8-GFP to chromatin during G1-exit and G2-exit meiosis. Centromere localization of Rec8-GFP at 2 h (upper panel) and association of Rec8-GFP with global chromatin at 4 h, shown by chromosome spread (lower panel). Scale bars, 10 μ m.

polymerase chain reaction (PCR) on the immunoprecipitate with specific primers, testing specifically the association of Rec8 with the centromeres. In fission yeast, the centromeres are organized into two distinct domains (Fig. 2d): an outer heterochromatic region consisting of highly repeated sequences, and an inner region consisting of unique sequences that interact with centromere-

specific proteins^{15,16}. Cytological observations indicated that Rec8 is localized at centromeres⁴ (Fig. 2c; 2 hours) and the CHIP assays established that Rec8 becomes associated with both the inner and outer centromere regions nearly 1 h before the pre-meiotic S phase (Fig. 2a, e; 2 hours). Rec8 became localized in the arm regions by 3 h as cells initiated the pre-meiotic S phase (Fig. 2e). Thus, Rec8 becomes associated with the chromatin found at the centromeres and chromosome arms before the pre-meiotic S phase.

We next examined whether *rec8* must be produced before the pre-meiotic S phase to generate reductional chromosome segregation during a normal meiosis. *Spo5* is expressed 1 h later than *rec8* (Fig. 1c). We therefore placed *rec8-GFP* (which functions similarly to *rec8*⁺) under the control of the *spo5* promoter; this [*spo5*-promoter]-*rec8-GFP* construct was then integrated at the chromosomal *rec8* locus (see Supplementary Information). Similarly, a [*rec8*-promoter]-*rec8-GFP* strain was constructed as a control. After the induction of G1-exit meiosis, pre-meiotic DNA replication occurred between 3 and 4 h and the first meiotic division occurred at 6 h in both strains (Fig. 3a, c). In [*rec8*-promoter]-*rec8-GFP* cells, the transcripts of *rec8-GFP* appeared before DNA replication, whereas in [*spo5*-promoter]-*rec8-GFP* cells the transcription of *rec8-GFP* appeared during DNA replication and thus the accumulation of Rec8-GFP protein in the nucleus was delayed until S/G2 (Fig. 3b, c). To assess whether chromosome segregation at meiosis I was reductional, we monitored the segregation of the centromere-linked markers *lys1* and *ade6* in the dyads formed. In [*rec8*-promoter]-*rec8-GFP* cells, the dyads exhibited a predominantly reductional segregation pattern for the centromere-linked markers (Fig. 3d). Notably, the dyads in [*spo5*-promoter]-*rec8-GFP* cells, which were at least 70% viable, demonstrated a predominantly equational segregation pattern (Fig. 3d). Furthermore, the recombination level between *lys1* and *leu2* loci was decreased nearly 20-fold (not shown). Thus, we conclude that the meiotic

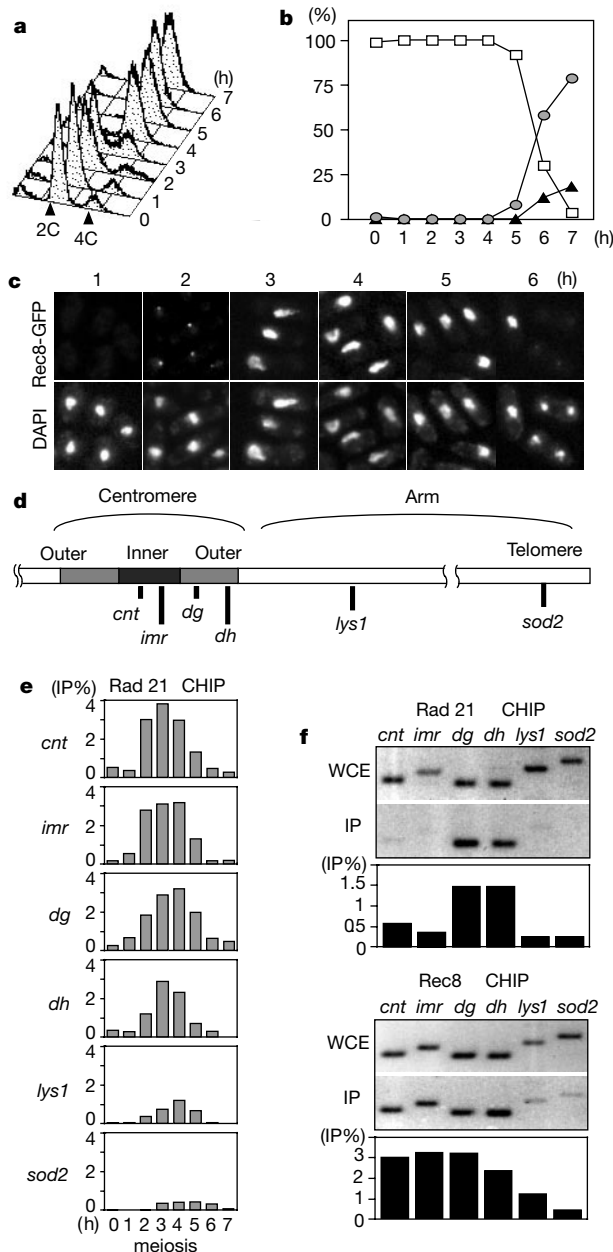


Figure 2 Association of Rec8 with centromeres and arms before the pre-meiotic S phase. Synchronous G1-exit meiosis was induced by using *rec8-GFP* cells and subjected to CHIP assays. **a**, Flow cytometric determination of the cellular DNA content. **b**, Number of nuclei per cell counted after staining with DAPI: squares, one nucleus; circles, two nuclei; triangles, three or four nuclei. **c**, Expression of Rec8-GFP (upper panels) and corresponding staining with DAPI (lower panels). **d**, Schematic representation of *Schizosaccharomyces pombe* chromosome I and the primers used (*cnt*, *imr*, *dg*, *dh*, *lys1*, *sod2*). **e**, Quantification of amplified PCR products from immunoprecipitates (IP) and the whole-cell extract, and calculation of the percentages of immunoprecipitated DNA. **f**, Analysis of the mitotically proliferating *rad21-GFP* cells (mostly in G2 phase) by CHIP. Representative PCRs of immunoprecipitates and the whole-cell extract (WCE) from *rad21-GFP* extracts (mitosis) and *rec8-GFP* extracts (at 4 h in meiosis) are shown. The percentage of immunoprecipitation is indicated at the bottom.

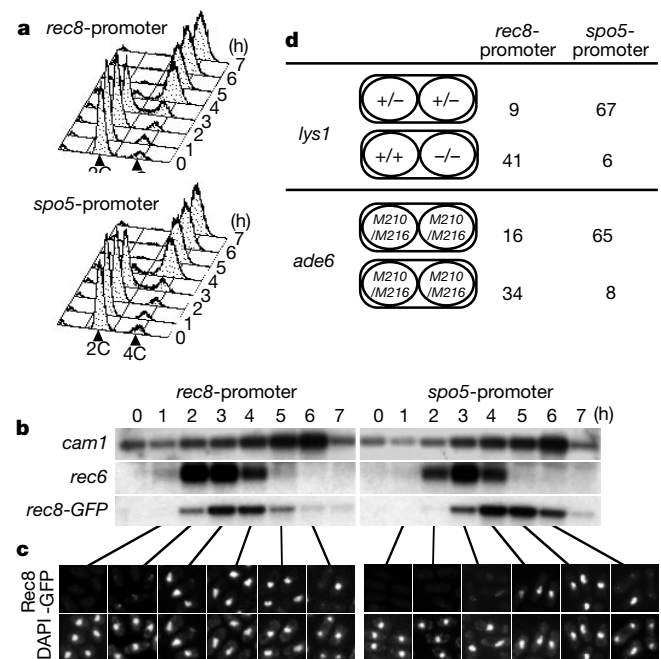


Figure 3 Meiotic cohesin Rec8 must be expressed before the pre-meiotic S phase for reductional division. The strains carrying [*rec8*-promoter]-*rec8-GFP* or [*spo5*-promoter]-*rec8-GFP* were analysed after G1-exit meiosis. **a**, Flow cytometric determination of the cellular DNA content. **b**, Transcripts of *rec8-GFP* analysed by Northern blotting; *rec6* and *cam1* were used as internal controls. **c**, Expression of Rec8-GFP (upper panels) and corresponding staining with DAPI (lower panels). **d**, Full viable dyads formed were examined for segregation of the centromere-linked markers.

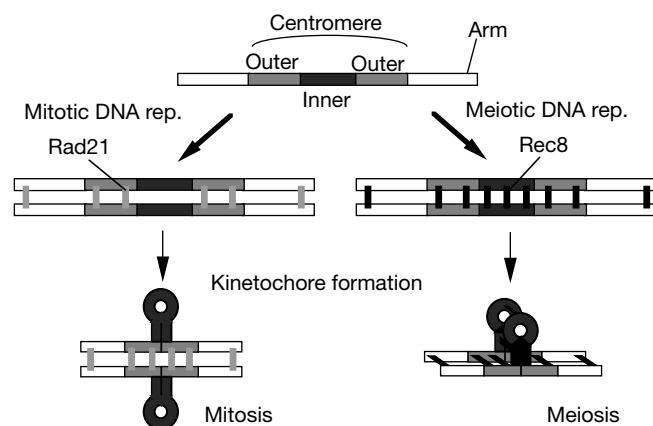


Figure 4 A model for the regulation of kinetochore orientation during meiosis by Rec8. The inner centromere regions are likely to form inverted configurations, protruding from the other regions of the centromeres¹⁵. Rec8 located at the inner centromere regions establishes cohesion during DNA replication, leading to monopolar orientation of the sister kinetochores at the subsequent nuclear division.

cohesin Rec8 must be expressed before the pre-meiotic S phase for meiosis I to become reductional in character and for a full level of recombination to occur.

To test whether the expression of Rec8 before S phase is sufficient to establish reductional division, we constructed cells expressing Rec8 under a constitutive promoter and examined for chromosome segregation during mitosis or G2-exit meiosis. In both cases, cells underwent chromosome segregation that was mostly equational (ref. 4, and data not shown). Therefore, although Rec8 has an essential role, one or more other meiosis-specific factors have to cooperate with Rec8 during DNA replication to establish reductional division. All *rec* genes that we examined are expressed before the onset of DNA replication (Fig. 1c) and are likely to function during the pre-meiotic S phase. The *rec11* gene, encoding a product similar to another cohesin subunit¹⁷, was also expressed during pre-meiotic G1/S (not shown) and might cooperate with Rec8 to bring about efficient reductional division.

Our results suggest strongly that DNA replication *de novo* is required for the Rec8-dependent processes to be established during meiosis. This is consistent with previous reports that mitotic cohesin complexes in budding yeast are established during DNA replication¹⁸ and that S phase takes longer in meiosis in the *rec8* mutant, implying some link of pre-meiotic DNA replication with the Rec8 function¹⁹. It has been shown in fission yeast that the inner centromere complexes are essential in establishing the bipolar orientation of sister centromeres during mitosis⁵. We have found that meiotic cohesin Rec8, but not mitotic cohesin Rad21, interacts strongly with the inner centromere region, although both cohesins interact with the outer centromere region (Fig. 2f). These results suggest that Rec8 might somehow modify the inner centromere structures to generate sister centromeres that become orientated to the same spindle pole. Because the pre-meiotic S phase is critical for this process, we propose that the location of Rec8 at the inner centromere might establish sister-strand cohesion during the replication of this region, leading to monopolar orientation of the sister kinetochores subsequently formed, which would otherwise be bipolar-oriented as in mitosis (Fig. 4). Rec8-dependent sister-arm cohesion would be also established by passage through the pre-meiotic S phase. Given that the *rec8* mutant fails to develop homologue-connecting structures^{14,20}, the absence of a pre-meiotic S phase would similarly decrease recombination. Taken together, our findings explain the importance of the pre-meiotic S phase for meiosis and provide an explanation of why meiosis is initiated from G1. The meiosis-specific cohesin Rec8 is conserved from yeast to the Metazoa, indicating that our conclusions are relevant for meiosis in all eukaryotes. □

Methods

Medium, strains and general methods

All media and growth conditions were as described²¹ unless otherwise stated. Two adenine-dependent mutations, *ade6-M26* and *ade6-469*, were used to monitor recombination¹¹. Other *ade6* alleles, *ade6-M210* and *ade6-M216*, exhibit inter-allelic complementation; homozygous *ade6-M210/ade6-M210* or *ade6-M216/ade6-M216* diploids form pink colonies on adenine-limiting (10 mg l⁻¹) plates, and heterozygous *ade6-M210/ade6-M216* diploids form white colonies. They were therefore used to monitor the segregation of the linked centromere in chromosome III. Spore analysis, GFP fluorescence and chromosome spread were as described previously⁴.

Synchronous G2-exit and G1-exit meioses

A *pat1^{ts} cdc2^{ts} h⁺/h⁺* diploid strain was cultured at its permissive temperature (25 °C) in the minimal medium EMM, and synchronous populations of early G2 cells were prepared by cell-size selection using centrifugal elutriation. Subsequent culturing of the cells at the nonpermissive temperature (34 °C) led to the inactivation of *pat1^{ts}*, thereby inducing ectopic meiosis (ref. 6; see also Supplementary Information). After 4 h they were shifted to 30 °C. The presence of the *cdc2^{ts}* mutation delayed progression through G2, so the experiment was not complicated by cells dividing and entering meiosis from G1 of the next cell cycle. This was confirmed by monitoring the cell number, which did not increase during meiosis. In this condition, 50–80% of the cells formed dyads because of the *cdc2^{ts}* mutation; the rest formed tetrads. For G1-exit meiosis, G1 cells were collected by elutriation after being cultured in the nitrogen-depleted medium EMM-N at 25 °C for 16 h. These cells were shifted to 34 °C and 0.5 g l⁻¹ of NH₄Cl was added; after 3 h, cultures were shifted to 30 °C. In Figs 2 and 3, G1 cells were collected by nitrogen depletion, but without elutriation, and shifted to 32 °C to induce meiosis.

CHIP assay

The procedure was done essentially as described previously²², with some modifications. We prepared extracts from [*rec8*-promoter]–*rec8*–GFP cells subjected to synchronous meiosis and immunoprecipitated by rabbit anti-GFP antibodies (Clontech). The linear range of the PCR reaction was determined by serial dilution of template DNA and all reactions were done within this range. PCR products were quantified with a Fuji Fluoro-Image Analyzer. We calculated the percentage of total DNA that was immunoprecipitated by comparing the intensity of PCR products from immunoprecipitated DNA. Asynchronous mitotic *rad21*–GFP cells, mostly at G2 phase, were examined similarly.

Received 7 August; accepted 10 November 2000.

- Kleckner, N. Meiosis: how could it work? *Proc. Natl. Acad. Sci. USA* **93**, 8167–8174 (1996).
- Roeder, G. S. Meiotic chromosomes: it takes two to tango. *Genes Dev.* **11**, 2600–2621 (1997).
- DeVaux, L. C. & Smith, G. R. Region-specific activators of meiotic recombination in *Schizosaccharomyces pombe*. *Genes Dev.* **8**, 203–210 (1994).
- Watanabe, Y. & Nurse, P. Cohesin Rec8 is required for reductional chromosome segregation at meiosis. *Nature* **400**, 461–464 (1999).
- Takahashi, K., Chen, E. S. & Yanagida, M. Requirement of Mis6 centromere connector for localizing a CENP-A-like protein in fission yeast. *Science* **288**, 2215–2219 (2000).
- Iino, Y. & Yamamoto, M. Negative control for the initiation of meiosis in *Schizosaccharomyces pombe*. *Proc. Natl. Acad. Sci. USA* **82**, 2447–2451 (1985).
- McLeod, M. & Beach, D. A specific inhibitor of the *ran1+* protein kinase regulates entry into meiosis in *Schizosaccharomyces pombe*. *Nature* **332**, 509–514 (1988).
- Watanabe, Y., Shinozaki-Yabana, S., Chikashige, Y., Hiraoka, Y. & Yamamoto, M. Phosphorylation of RNA-binding protein controls cell cycle switch from mitotic to meiotic in fission yeast. *Nature* **386**, 187–190 (1997).
- Yamamoto, M., Imai, Y. & Watanabe, Y. in *The Molecular and Cellular Biology of the Yeast Saccharomyces* (eds Pringle, J., Broach, J. & Jones, E.) 1037–1106 (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1997).
- Horie, S. *et al.* The *Schizosaccharomyces pombe mei4+* gene encodes a meiosis-specific transcription factor containing a forkhead DNA-binding domain. *Mol. Cell. Biol.* **18**, 2118–2129 (1998).
- Bähler, J., Schuchert, P., Grimm, C. & Kohli, J. Synchronized meiosis and recombination in fission yeast: observations with *pat1-114* diploid cells. *Curr. Genet.* **19**, 445–451 (1991).
- Niwa, O. & Yanagida, M. Universal and essential role of MPF/cdc2+. *Nature* **336**, 430 (1988).
- Nabeshima, K. *et al.* Dynamics of centromeres during metaphase–anaphase transition in fission yeast: Dis1 is implicated in force balance in metaphase bipolar spindle. *Mol. Biol. Cell* **9**, 3211–3225 (1998).
- Klein, F. *et al.* A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. *Cell* **98**, 91–103 (1999).
- Takahashi, K. *et al.* A low copy number central sequence with strict symmetry and unusual chromatin structure in fission yeast centromere. *Mol. Biol. Cell* **3**, 819–835 (1992).
- Partridge, J. F., Borgstrom, B. & Allshire, J. C. Distinct protein interaction domains and protein spreading in a complex centromere. *Genes Dev.* **14**, 783–791 (2000).
- Toth, A. *et al.* Yeast cohesin complex requires a conserved protein, Eco1p (Ctf7), to establish cohesion between sister chromatids during DNA replication. *Genes Dev.* **13**, 320–333 (1999).
- Uhlmann, F. & Nasmyth, K. Cohesion between sister chromatids must be established during DNA replication. *Curr. Biol.* **8**, 1095–1101 (1998).
- Cha, R. S., Weiner, B. M., Keeney, S., Dekker, J. & Kleckner, N. Progression of meiotic DNA replication is modulated by interchromosomal interaction proteins, negatively by Spo11p and positively by Rec8p. *Genes Dev.* **14**, 493–503 (2000).
- Molnar, M., Bähler, J., Sipiczki, M. & Kohli, J. The *rec8* gene of *Schizosaccharomyces pombe* is involved in linear element formation, chromosome pairing and sister-chromatid cohesion during meiosis. *Genetics* **141**, 61–73 (1995).
- Moreno, S., Klar, A. & Nurse, P. Molecular genetic analysis of fission yeast *Schizosaccharomyces pombe*. *Methods Enzymol.* **194**, 795–823 (1991).

22. Saitoh, S., Takahashi, K. & Yanagida, M. Mis6, a fission yeast inner centromere protein, acts during G1/S and forms specialized chromatin required for equal segregation. *Cell* **90**, 131–143 (1997).
23. Kelly, T. J. *et al.* The fission yeast *cdc18* gene product couples S-phase to start and mitosis. *Cell* **74**, 371–382 (1993).
24. Fernandez-Sarabia, M. J., McNery, C., Harris, P., Gordon, C. & Fantes, P. The cell cycle genes *cdc22⁺* and *suc22⁺* of the fission yeast *Schizosaccharomyces pombe* encode the large and small subunits of ribonucleotide reductase. *Mol. Gen. Genet.* **238**, 241–251 (1993).
25. Lin, L. & Smith, G. R. Transient, meiosis-induced expression of the *rec6* and *rec12* genes of *Schizosaccharomyces pombe*. *Genetics* **136**, 769–779 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. P. Cooper for critical reading of the manuscript and M. Yanagida for CHIP method. Y.W. thanks all the members of P.N.'s laboratory for help and discussion, particularly J. Hayles, H. Murakami and G. Simchen. Y.W. was supported by JSPS and Uehara fellowships and grants from the Ministry of Education, Science and Culture of Japan.

Correspondence and requests for materials should be addressed to Y.W. (e-mail: ywatanab@ims.u-tokyo.ac.jp).

Role for a bidentate ribonuclease in the initiation step of RNA interference

Emily Bernstein^{*†}, Amy A. Caudy^{*‡}, Scott M. Hammond^{*§} & Gregory J. Hannon^{*}

^{*} Cold Spring Harbor Laboratory, and [‡]Watson School of Biological Sciences, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

[†] Graduate Program in Genetics, State University of New York at Stony Brook, Stony Brook, New York, 11794, USA

[§] Genetics, 1 Kendall Square, Building 600, Cambridge, Massachusetts 01239, USA

RNA interference (RNAi) is the mechanism through which double-stranded RNAs silence cognate genes^{1–5}. In plants, this can occur at both the transcriptional and the post-transcriptional levels^{1,2,5}; however, in animals, only post-transcriptional RNAi has been reported to date. In both plants and animals, RNAi is characterized by the presence of RNAs of about 22 nucleotides in length that are homologous to the gene that is being suppressed^{6–8}. These 22-nucleotide sequences serve as guide sequences that instruct a multicomponent nuclease, RISC, to destroy specific messenger RNAs⁶. Here we identify an enzyme, Dicer, which can produce putative guide RNAs. Dicer is a member of the RNase III family of nucleases that specifically cleave double-stranded RNAs, and is evolutionarily conserved in worms, flies, plants, fungi and mammals. The enzyme has a distinctive structure, which includes a helicase domain and dual RNase III motifs. Dicer also contains a region of homology to the RDE1/QDE2/ARGONAUTE family that has been genetically linked to RNAi^{9,10}.

Biochemical studies have suggested that post-transcriptional gene silencing (PTGS) is accomplished by a multicomponent nuclease that targets mRNAs for degradation^{6,8,11}. The specificity of this complex may derive from the incorporation of a small guide sequence that is homologous to the mRNA substrate⁶. These ~22-nucleotide RNAs, originally identified in plants that were actively silencing transgenes⁷, have been produced during RNAi *in vitro* using an extract prepared from *Drosophila* embryos⁸. Putative guide RNAs can also be produced in extracts from *Drosophila* S2 cells (Fig. 1a). To investigate the mechanism of PTGS, we have performed both biochemical fractionation and candidate gene approaches to identify the enzymes that execute each step of RNAi.

Our previous studies resulted in the partial purification of an enzyme complex, RISC, which is an effector nuclease for RNA interference⁶. This enzyme was isolated from *Drosophila* S2 cells in which RNAi had been initiated *in vivo* by transfection with double-stranded RNA (dsRNA). We first investigated whether the RISC enzyme, and the enzyme that initiates RNAi through processing of dsRNA into 22-nucleotide sequences, are distinct activities. RISC activity could be largely cleared from extracts by high-speed centrifugation (100,000g for 60 min), whereas the activity that produces 22-nucleotide sequences remained in the supernatant (Fig. 1b, c). This simple fractionation indicates that RISC and the 22-nucleotide sequence-generating activity may be separable. However, it seems probable that these enzymes interact at some point during the silencing process, and it remains possible that initiator and effector enzymes share common subunits.

RNase III family members are among the few nucleases that show specificity for dsRNA¹². Analysis of the *Drosophila* and *Caenorhabditis elegans* genomes reveals several types of RNase III enzymes. First is the canonical RNase III, which contains a single RNase III signature motif and a dsRNA-binding domain (dsRBD; for example RNC_CAEEL). Second is a class represented by Drosha¹³, a *Drosophila* enzyme that contains two RNase III motifs and a dsRBD (CeDrosha in *C. elegans*). A third class contains two RNase III signatures and an amino-terminal helicase domain (for example, *Drosophila* CG4792 and CG6493; *C. elegans* K12H4.8), which had been proposed as potential RNAi nucleases^{14,20}. We tested representatives of all three classes for the ability to produce discrete RNAs of ~22 nucleotides from dsRNA substrates.

To test the dual RNase III enzymes, we prepared variants of Drosha and CG4792 tagged with the T7 epitope. These were expressed in transfected S2 cells and isolated by immunoprecipitation using antibody–agarose conjugates. Treatment of the dsRNA with the CG4792 immunoprecipitate yielded fragments of about 22 nucleotides, similar to those produced in either the S2 or embryo extracts (Fig. 2a). Neither the activity in extract nor that in immunoprecipitates depended on the sequence of the RNA substrate, as dsRNAs derived from several genes were processed equivalently (see Supplementary Information). Negative results

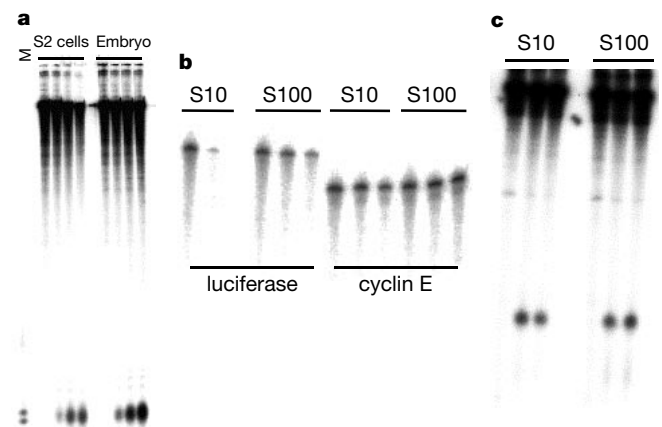


Figure 1 Generation of 22-nucleotide sequences and degradation of mRNA by distinct enzymatic complexes. **a**, Extracts prepared from 0–12 h *Drosophila* embryos or *Drosophila* S2 cells. Extracts were incubated for 0, 15, 30 or 60 min (left to right) with a uniformly labelled dsRNA. A 22-nucleotide marker prepared by *in vitro* transcription of a synthetic template is indicated (M). **b**, Whole-cell extracts from S2 cells transfected with luciferase dsRNA. S10 represents our standard RISC extract⁶. S100 extracts were prepared by additional centrifugation of S10 extracts for 60 min at 100,000g. Assays for mRNA degradation⁶ were performed for 0, 30 or 60 min (left to right in each set) with either a single-stranded luciferase mRNA or a single-stranded cyclin E mRNA, as indicated. **c**, S10 or S100 extracts incubated with cyclin E dsRNAs for 0, 60 or 120 min (left to right).

were obtained with Drosha and with immunoprecipitates of a DExH box helicase (Homeless¹⁵; see Fig. 2a and b). Western blotting confirmed that each of the tagged proteins was expressed and immunoprecipitated similarly (see Supplementary Information). Thus, we conclude that CG4792 may carry out the initiation step of RNAi by producing guide sequences of about 22 nucleotides from dsRNAs. Because of its ability to digest dsRNA into uniformly sized, small RNAs, we have named this enzyme Dicer (*Dcr*). *Dicer* mRNA is expressed in embryos, in S2 cells and in adult flies, which is consistent with the presence of functional RNAi machinery in all of these contexts (see Supplementary Information).

An antiserum directed against the carboxy terminus of the Dicer protein (Dicer-1, CG4792) could immunoprecipitate a nuclease activity from either the *Drosophila* embryo extracts or from S2 cell lysates that produced RNAs of about 22 nucleotides from dsRNA substrates (Fig. 2c). The putative guide RNAs that are produced by the Dicer-1 enzyme precisely co-migrate with 22-nucleotide sequences that are produced in extract, and with 22-nucleotide sequences that are associated with the RISC enzyme (Fig. 2d, f). The enzyme that produces guide RNAs in *Drosophila* embryo extracts is ATP dependent⁸. Depletion of this cofactor resulted in a roughly sixfold reduction of dsRNA cleavage rate and in the production of

RNAs with a slightly lower mobility. Of note, both Dicer-1 immunoprecipitates and extracts from S2 cells require ATP for the production of ~22-nucleotide sequences (Fig. 2d). We did not observe the accumulation of lower-mobility products in these cases, although we did routinely observe these in ATP-depleted embryo extracts. The requirement of this nuclease for ATP is an unusual property, and may indicate that unwinding of guide RNAs by the helicase domain is required for the enzyme to act catalytically.

For efficient induction of RNAi in *C. elegans* and in *Drosophila*, the initiating RNA must be double-stranded and must also be several hundred nucleotides in length⁴. Similarly, Dicer was inactive against single-stranded RNAs regardless of length (see Supplementary Information). The enzyme could digest both 200- and 500-nucleotide dsRNAs, but was significantly less active with shorter substrates (see Supplementary Information). In contrast, *Escherichia coli* RNase III could digest to completion dsRNAs of 35 or 22 nucleotides (data not shown). This suggests that the substrate preferences of the Dicer enzyme may contribute to, but not wholly determine, the size dependence of RNAi.

To determine whether the Dicer enzyme is involved in RNAi *in vivo*, we depleted Dicer activity from S2 cells and tested the effect on dsRNA-induced gene silencing. Transfection of S2 cells with a

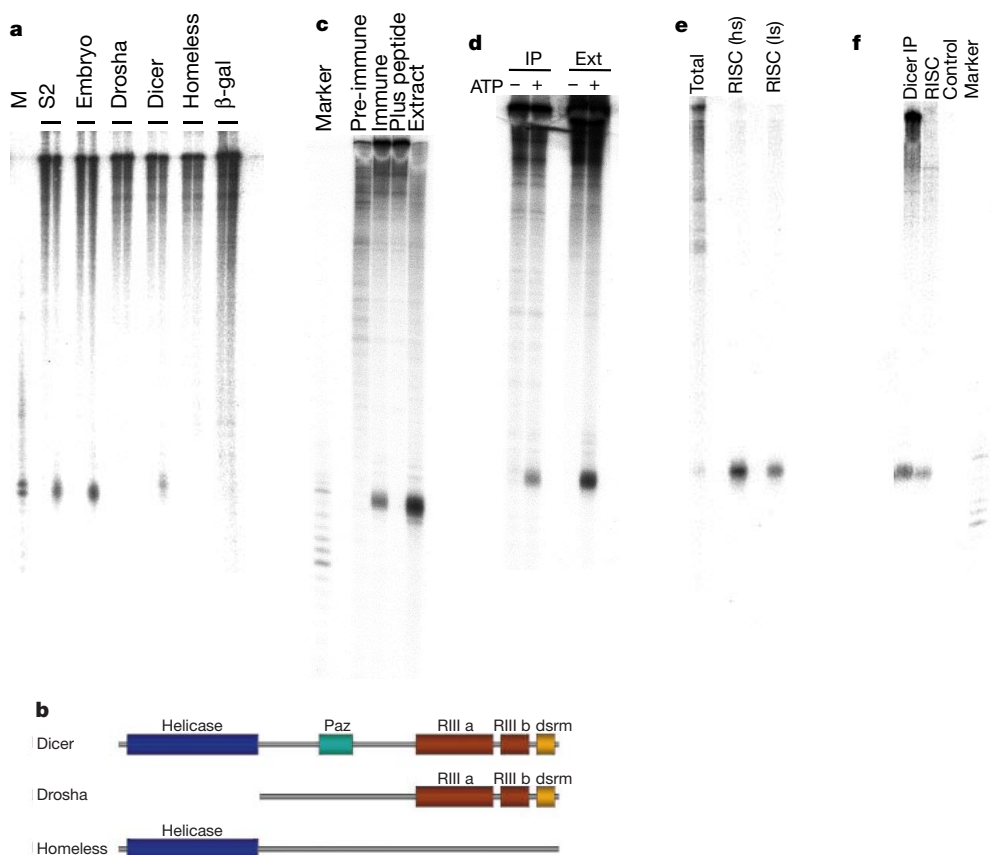


Figure 2 Production of 22-nucleotide sequences by CG4792/Dicer. **a**, *Drosophila* S2 cells transfected with plasmids that direct expression of T7-epitope-tagged versions of Drosha, CG4792/Dicer-1 and Homeless or untagged β -galactosidase. Proteins were immunoprecipitated and incubated with *cyclin E* dsRNA for 0 or 60 min. Reactions in *Drosophila* embryo and S2 cell extracts are shown. **b**, Domain structures of CG4792/Dicer-1, Drosha and Homeless. **c**, Immunoprecipitates prepared from detergent lysates of S2 cells using Dicer antiserum. As controls, similar preparations were made with a pre-immune serum and an immune serum that had been pre-incubated with an excess of antigenic peptide. Cleavage reactions in which each of these precipitates was incubated with a ~500 nucleotide fragment of *Drosophila* cyclin E are shown. An incubation of the substrate in *Drosophila* embryo extract is shown. **d**, Dicer immunoprecipitates incubated with dsRNA substrates in presence or absence of ATP. The same substrate was also

incubated with ATP-added or ATP-depleted S2 extracts. **e**, *Drosophila* S2 cells transfected with uniformly, ³²P-labelled dsRNA corresponding to the first 500 nucleotides of GFP. RISC complex was affinity purified using a histidine-tagged version of *Drosophila* Ago-2, a component of the RISC complex (Hammond *et al.*, manuscript in preparation). RISC was isolated under ribosome-associated (ls, low salt) or soluble, ribosome-extracted (hs, high salt) conditions⁶. The spectrum of labelled RNAs in the total lysate is shown. **f**, Comparison of guide RNAs produced by incubation of dsRNA with a Dicer immunoprecipitate, with guide RNAs present in affinity-purified RISC complex. These co-migrate on a gel that has single-nucleotide resolution. The control lane shows an affinity selection for RISC from cells transfected with labelled dsRNA, but not with the epitope-tagged *Drosophila* Ago-2.

mixture of dsRNAs homologous to the two *Drosophila* Dicer genes (CG4792 and CG6493) resulted in a roughly 6–7-fold reduction of Dicer activity either in whole-cell lysates or in Dicer-1 immunoprecipitates (Fig. 3a and b). Transfection with a control dsRNA (murine caspase-9) had no effect. Qualitatively similar results were seen if Dicer mRNA was examined by northern blotting (data not shown). Depletion of Dicer substantially compromised the ability of cells to silence an exogenous, green fluorescent protein (GFP) transgene by RNAi (Fig. 3c). These results indicate that Dicer may be involved in RNAi *in vivo*. The lack of complete inhibition of silencing may result from an incomplete suppression of Dicer or may indicate that *in vivo* guide RNAs may be produced by more than one mechanism.

Our results indicate that the process of RNAi can be divided into at least two distinct steps. Initiation of PTGS would occur on processing of a dsRNA by Dicer into ~22-nucleotide guide sequences, although we cannot formally exclude the possibility that another Dicer-associated nuclease may participate in this process. These guide RNAs would be incorporated into a distinct nuclease complex (RISC) that targets single-stranded mRNAs for degradation. An implication of this model is that the guide sequences are themselves derived directly from the dsRNA that triggers the response. In accord with this model, we have shown that ³²P-labelled, exogenous dsRNAs that have been introduced into S2 cells by transfection are incorporated into the RISC enzyme as 22-nucleotide sequences (Fig. 2e).

A notable feature of the Dicer family is its evolutionary conservation. Homologues are found in *C. elegans* (K12H4.8), *Arabidopsis* (for example, CARPEL FACTORY¹⁶, T25K16.4 and AC012328_1), mammals (Helicase-MOI¹⁷) and *Schizosaccharomyces pombe* (YC9A_SCHPO) (see Supplementary Information for comparisons). In fact, the human Dicer family member is capable of generating ~22-nucleotide RNAs from dsRNA substrates (see

Supplementary Information), which indicates that these structurally similar proteins may all share similar biochemical functions. Exogenous dsRNAs can affect gene function in early mouse embryos¹⁸, and our results suggest that this regulation may be accomplished by evolutionarily conserved RNAi machinery.

In addition to RNase III and helicase motifs, searches of the PFAM database indicate that each Dicer family member also contains a PAZ domain (see Supplementary Information)^{19,20}. This sequence was defined on the basis of its conservation in the Zwiile/ARGONAUTE/Piwi family that has been implicated in RNAi by mutations in *C. elegans* (Rde-1)⁹ and *Neurospora* (Qde-2)¹⁰. Although the function of this domain is unknown, it is notable that this region of homology is restricted to two gene families that participate in dsRNA-dependent silencing. Both the ARGONAUTE and Dicer families have also been implicated in common biological processes, namely the determination of stem-cell fates. A hypomorphic allele of *carpel factory*, a member of the Dicer family in *Arabidopsis*, is characterized by increased proliferation in floral meristems¹⁶. This phenotype and a number of other characteristic features are also shared by *Arabidopsis* ARGONAUTE (*ago1-1*) mutants²¹ (C. Kidner and R. Martienssen, personal communication). These genetic analyses provide evidence that RNAi may be more than a defensive response to unusual RNAs, but may also have integral functions in the regulation of endogenous genes.

With the identification of Dicer as a potential catalyst of the initiation step of RNAi, we have begun to unravel the biochemical basis of this unusual mechanism of gene regulation. It is now important to determine whether the conserved family members from other organisms, particularly mammals, also have a function in dsRNA-mediated gene regulation.

Note added in proof: Yang *et al.*²² have recently presented evidence that guide RNAs are derived directly from dsRNA in *Drosophila* embryos. Fagard *et al.*²³ have recently shown that *Arabidopsis* Ago1 is involved in PTGS. □

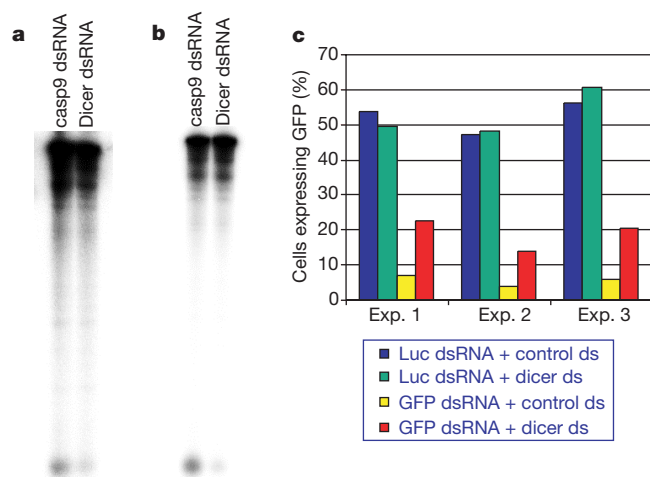


Figure 3 Dicer participates in RNAi. **a**, *Drosophila* S2 cells transfected with dsRNAs corresponding to the two *Drosophila* Dicers (CG4792 and CG6493) or control dsRNA corresponding to murine caspase-9 (casp9). Cytoplasmic extracts of these cells were tested for Dicer activity. Transfection with Dicer dsRNA reduces activity in lysates 7.4-fold. **b**, Dicer-1 antiserum (CG4792) used to prepare immunoprecipitates from S2 cells (treated as above). Dicer dsRNA reduces the activity of Dicer-1 6.2-fold. **c**, GFP expression of co-transfected cells. Three independent experiments were quantified by FACS. A comparison of the relative percentage of GFP-positive cells is shown for control (GFP plasmid plus luciferase dsRNA) or silenced (GFP plasmids plus GFP dsRNA) populations in cells that had previously been transfected with either control (caspase-9) or Dicer dsRNAs.

Methods

Plasmid constructs

A full-length complementary DNA encoding Drosha was obtained by polymerase chain reaction (PCR) from an expressed sequence tag sequenced by the Berkeley *Drosophila* genome project. The T7 epitope-tag was added to the N terminus of each cDNA by PCR, and the tagged cDNAs were cloned into pRIP—a retroviral vector designed specifically for expression in insect cells (E. B., unpublished observations). In this vector, expression is driven by the *Orygia pseudotsugata* IE2 promoter (Invitrogen). As no cDNA was available for CG4792/Dicer, a genomic clone was amplified from a bac (bacterial artificial chromosome) (BACR23F10; obtained from the BACPAC Resource Center in the Department of Human Genetics at the Roswell Park Cancer Institute). We added a T7 epitope tag at the N terminus of the coding sequence during amplification. We isolated the human Dicer gene from a cDNA library prepared from HaCaT cells (G.J.H., unpublished observations). A T7-tagged version of the complete coding sequence was cloned into pCDNA3 (Invitrogen) for expression in human cells (LinX-A).

Cell culture and extract preparation

We cultured S2 cells at 27 °C in 5% CO₂ in Schneider's insect media supplemented with 10% heat-inactivated fetal bovine serum (Gemini) and 1% antibiotic–antimycotic solution (Gibco BRL). Cells were collected for extract preparation at 10⁷ cells per ml. The cells were washed in PBS and resuspended in a hypotonic buffer (10 mM HEPES pH 7.0, 2 mM MgCl₂ and 6 mM β-mercaptoethanol) and lysed. We centrifuged cell lysates at 20,000g for 20 min. We stored extracts at –80 °C. We reared *Drosophila* embryos in fly cages by standard methodologies and collected them every 12 h. We dechorionated the embryos in 50% chlorox bleach and washed them thoroughly with distilled water. Lysis buffer (10 mM Hepes, 10 mM KCl, 1.5 mM MgCl₂, 0.5 mM EGTA, 10 mM β-glycerophosphate, 1 mM dithiothreitol (DTT) and 0.2 mM PMSF) was added to the embryos, and extracts were prepared by homogenization in a tissue grinder. Lysates were centrifuged for 2 h at 200,000g, and were frozen at –80 °C. LinX-A cells, a highly transfectable derivative of human 293 cells (L. Xie and G.J.H., unpublished observations) were maintained in DMEM/10% FCS.

Transfections and immunoprecipitations

We transfected S2 cells using a calcium phosphate procedure essentially as described⁶. Transfection rates were about 90%, as monitored in controls using an *in situ* β-galactosidase assay. We also transfected LinX-A cells by calcium phosphate co-precipitation. For immunoprecipitations, cells (~5 × 10⁶ per immunoprecipitate) were

transfected with various clones, and lysed 3 d later in immunoprecipitation buffer (125 mM KOAc, 1 mM MgOAc, 1 mM CaCl₂, 5 mM EGTA, 20 mM HEPES pH 7.0, 1 mM DTT and 1% Nonidet P40 plus complete protease inhibitors (Roche)). We centrifuged lysates for 10 min at 14,000g, and then added supernatants to T7 antibody-agarose beads (Novagen). We performed antibody binding for 4 h at 4 °C. Beads were centrifuged and washed three times in lysis buffer, and once in reaction buffer. The Dicer antiserum was raised in rabbits using a keyhole limpet haemocyanin-conjugated peptide corresponding to the C-terminal eight amino acids of *Drosophila* Dicer-1 (CG4792).

Cleavage reactions

Templates to be transcribed to dsRNA were generated by PCR with forward and reverse primers, each containing a T7 promoter sequence. RNAs were produced using Riboprobe kits (Promega) and were uniformly labelled during the transcription reaction with ³²P-labelled UTP. Single-stranded RNAs were purified from 1% agarose gels. For cleavage of dsRNA, 5 µl of embryo or S2 extracts were incubated for 1 h at 30 °C with dsRNA in a reaction containing 20 mM HEPES pH 7.0, 2 mM magnesium acetate, 2 mM DTT, 1 mM ATP and 5% Supersin (Ambion). Immunoprecipitates were treated similarly, except that a minimal volume of reaction buffer (including ATP and Supersin) and dsRNA were added to beads that had been washed in reaction buffer. For ATP depletion, *Drosophila* embryo extracts were incubated for 20 min at 30 °C with 2 mM glucose and 0.375 U of hexokinase (Roche), before the addition of dsRNA.

Northern and western analysis

Total RNA was prepared from *Drosophila* embryos (0–12 h), from adult flies and from S2 cells using Trizol (Lifetech). We isolated mRNA by affinity selection using magnetic LIGO-dT beads (Dyna). RNAs were electrophoresed on denaturing formaldehyde/agarose gels, blotted and probed with randomly primed DNAs corresponding to Dicer. For western analysis, T7-tagged proteins were immunoprecipitated from whole-cell lysates in immunoprecipitation buffer using agarose-conjugated anti-T7 antibody. Proteins were released from the beads by boiling in Laemmli buffer, and were separated by 8% SDS-polyacrylamide gel electrophoresis. After transfer to nitrocellulose, proteins were visualized using an HRP-conjugated anti-T7 antibody (Novagen) and chemiluminescent detection (Supersignal, Pierce).

RNAi of Dicer

Drosophila S2 cells were transfected either with a dsRNA corresponding to mouse caspase-9 or with a mixture of two dsRNAs corresponding to *Drosophila* Dicer-1 and Dicer-2 (CG4792 and CG6493). Two days after the initial transfection, cells were again transfected with a mixture containing a GFP expression plasmid and either luciferase dsRNA or GFP dsRNA as described⁶. Cells were assayed for Dicer activity or fluorescence 3 d after the second transfection. Quantification of fluorescent cells was done on a Coulter EPICS cell sorter, after fixation. Control transfections indicated that Dicer activity was not affected by the introduction of caspase-9 dsRNA.

Received 16 October; accepted 14 November 2000.

1. Baulcombe, D. C. RNA as a target and an initiator of post-transcriptional gene silencing in transgenic plants. *Plant Mol. Biol.* **32**, 79–88 (1996).
2. Wassenaar, M. & Pelissier, T. A model for RNA-mediated gene silencing in higher plants. *Plant Mol. Biol.* **37**, 349–62 (1998).
3. Montgomery, M. K. & Fire, A. Double-stranded RNA as a mediator in sequence-specific genetic silencing and co-suppression. *Trends Genet.* **14**, 255–258 (1998).
4. Sharp, P. A. RNAi and double-strand RNA. *Genes Dev.* **13**, 139–141 (1999).
5. Sijen, T. & Kooter, J. M. Post-transcriptional gene-silencing: RNAs on the attack or on the defense? *BioEssays* **22**, 520–531 (2000).
6. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
7. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952 (1999).
8. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
9. Tabara, H. *et al.* The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**, 123–132 (1999).
10. Catalanotto, C., Azzalin, G., Macino, G. & Cogoni, C. Gene silencing in worms and fungi. *Nature* **404**, 245 (2000).
11. Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P. & Sharp, P. A. Targeted mRNA degradation by double-stranded RNA in *Drosophila* cells. *Genes Dev.* **13**, 3191–3197 (1999).
12. Nicholson, A. W. Function, mechanism and regulation of bacterial ribonucleases. *FEMS Microbiol. Rev.* **23**, 371–390 (1999).
13. Filippov, V., Solov'yev, V., Filipova, M. & Gill, S. S. A novel type of RNase III family proteins in eukaryotes. *Gene* **245**, 213–221 (2000).
14. Bass, B. L. Double-stranded RNA as a template for gene silencing. *Cell* **101**, 235–238 (2000).
15. Gillespie, D. E. & Berg, C. A. Homeless is required for RNA localization in *Drosophila* oogenesis and encodes a new member of the DE-H family of RNA-dependent ATPases. *Genes Dev.* **9**, 2495–2508 (1995).
16. Jacobsen, S. E., Running, M. P. & Meyerowitz, E. M. Disruption of an RNA helicase/RNase III gene in *Arabidopsis* causes unregulated cell division in floral meristems. *Development* **126**, 5231–5243 (1999).
17. Matsuda, S. *et al.* Molecular cloning and characterization of a novel human gene (HERNA) which encodes a putative RNA-helicase. *Biochim. Biophys. Acta* **1490**, 163–169 (2000).
18. Wianny, F. & Zernicka-Goetz, M. Specific interference with gene function by double-stranded RNA in early mouse development. *Nature Cell Biol.* **2**, 70–75 (2000).
19. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Protein Struct. Funct. Genet.* **28**, 405–420 (1997).

20. Cerutti, L., Mian, N. & Bateman, A. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* **25**, 481–482 (2000).
21. Bohmert, K. *et al.* AGO1 defines a novel locus of *Arabidopsis* controlling leaf development. *EMBO J.* **17**, 170–180 (1998).
22. Yang, D., Lu, H. & Erickson, J. W. Evidence that processed small dsRNAs may mediate sequence-specific mRNA degradation during RNAi in *Drosophila* embryos. *Curr. Biol.* **10**, 1191–1200 (2000).
23. Fagard, M., Bouter, S., Morel, J. B., Bellini, C. & Vaucheret, H. AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl Acad. Sci. USA* **97**, 11650–11654 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank A. Nicholson for his gift of purified RNase III, and P. Fisher, M. McConnel and M. Pang for providing aid and materials for large-scale fly embryo culture. The *Homeless* clone was a gift from D. E. Gillespie and C. A. Berg. We also thank R. Kobayashi and R. Martienssen for discussion and critical reading of the manuscript, and K. Velinon for FACS. A.A.C. is an Anderson Fellow of the Watson School of Biological Sciences and a Predoctoral Fellow of the Howard Hughes Medical Institute. S.M.H. is a visiting scientist from Genetics, (Cambridge, MA). G.J.H. is a Pew Scholar in the biomedical sciences. This work was supported in part by grants from the NIH (G.J.H.).

A model for SOS-lesion-targeted mutations in *Escherichia coli*

Phuong Pham*, Jeffrey G. Bertram*, Mike O'Donnell†, Roger Woodgate‡ & Myron F. Goodman*

* Department of Biological Sciences and Chemistry, Hedco Molecular Biology Laboratories, University of Southern California, University Park, Los Angeles, California 90089-1340, USA

† Rockefeller University and Howard Hughes Medical Institute, New York, New York 10021, USA

‡ Section on DNA Replication, Repair and Mutagenesis, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892-2725, USA

The UmuD'2C protein complex (*Escherichia coli* pol V)^{1–3} is a low-fidelity DNA polymerase (pol) that copies damaged DNA in the presence of RecA, single-stranded-DNA binding protein (SSB) and the β,γ-processivity complex of *E. coli* pol III (ref. 4). Here we propose a model to explain SOS-lesion-targeted mutagenesis, assigning specific biochemical functions for each protein during translesion synthesis. (SOS lesion-targeted mutagenesis occurs when pol V is induced as part of the SOS response to DNA damage and incorrectly incorporates nucleotides opposite template lesions.) Pol V plus SSB catalyses RecA filament disassembly in the 3' to 5' direction on the template, ahead of the polymerase, in a reaction that does not involve ATP hydrolysis. Concurrent ATP-hydrolysis-driven filament disassembly in the 5' to 3' direction results in a bidirectional stripping of RecA from the template strand. The bidirectional collapse of the RecA filament restricts DNA synthesis by pol V to template sites that are proximal to the lesion, thereby minimizing the occurrence of untargued mutations at undamaged template sites.

Lesions that block DNA replication persist in prokaryotic and eukaryotic cells despite the presence of base excision, nucleotide excision and postreplication repair⁵. A group of DNA polymerases have been discovered (the UmuC/DinB/Rad30/Rev1 superfamily) whose function is to copy DNA template lesions⁶. The presence of *E. coli* pol V (UmuD'2C) is essential for SOS-induced mutagenesis⁵. However, pol V cannot catalyse translesion synthesis (TLS) by itself; it requires the presence of RecA and single-stranded-DNA binding protein (SSB), and is stimulated by β-sliding clamp in a 'mutasomal' complex⁷ (pol V Mut) to copy replication-blocking lesions⁴.

We have reconstituted TLS catalysed by pol V Mut *in vitro* using synthetic DNA oligomers to determine the function of each mutasomal component during TLS.

The effect of using different combinations of RecA, SSB and β,γ -complex on pol V activity was analysed on damaged (abasic moiety, X) and undamaged primer/template DNA, which consisted of a 30-nucleotide primer and a 240-nucleotide DNA template (Fig. 1). In contrast with assembly of a RecA filament—assembly in a 5' to 3' direction on single-stranded DNA, either with ATP or with its poorly hydrolysable analogue ATP- γ S—filament disassembly, in the same direction, requires ATP hydrolysis⁸. We have named the filaments formed in the presence of ATP- γ S 'stabilized' RecA filaments.

Primer extension by means of nucleotide insertion opposite template G is very slow for pol V alone, that is, in the absence of mutasomal accessory proteins, with or without DNA damage (Fig. 1a and b, lane 1). Addition of SSB in the absence of RecA strongly stimulates extension (Fig. 1a and b, lane 2), and addition of SSB plus β,γ -complex facilitates further synthesis to the end of the undamaged template strand (Fig. 1b, lane 4). Addition of RecA plus ATP causes a marked stimulation of synthesis (Fig. 1a and b, compare lanes 1 and 5), whereas significant TLS opposite X and beyond occurs only in the presence of pol V Mut, for example, pol V, RecA, SSB and β,γ -complex (Fig. 1a, lane 8).

The average processivity of pol V Mut synthesis increases by at least 3–5-fold when ATP is replaced by ATP- γ S, from 12 nucleotides per template-binding event (Fig. 1b, lane 8) to >38 nucleotides (Fig. 1b, lane 12) on the undamaged template, and from 7 to >36 nucleotides on the lesion-containing template (Fig. 1a, lanes 8 and 12, respectively). The fraction of full-length product to total DNA synthesis changes more markedly for RecA filaments formed in the presence of ATP- γ S. Only 2% of the product is full length in the presence of ATP (Fig. 1b, lane 8) compared with 50% using ATP- γ S (Fig. 1b, lane 12). This difference is more pronounced on the lesion-containing template where less than 0.1% of the primers extended to the end of the template with ATP (Fig. 1a, lane 8) compared with 50% with ATP- γ S (Fig. 1a, lane 12). Synthesis is less processive when β,γ -complex is omitted from the reaction. The average processivity

on undamaged DNA is reduced from >38 to 15 nucleotides (Fig. 1b, lane 10), and from >36 to 4 nucleotides on damaged DNA (Fig. 1a, lane 10). The fraction of primers that extend to the end of the template diminishes from 51% to 4% in the absence of β,γ -complex on undamaged DNA, and from 49% to <0.1% on the lesion-containing template.

Notably, no detectable synthesis occurs in the absence of SSB when a stabilized RecA filament is formed on either of the templates (Fig. 1a and b, lane 11). Relatively strong synthesis occurs in the absence of SSB when the filament is formed with ATP, but the synthesis is essentially non-processive even in the presence of β,γ -complex (Fig. 1a and b, lane 7). The stabilized filament cannot be copied when pol III replaces pol V in the assay (Fig. 1c, lanes 10–12), whereas pol III is able to replicate a RecA filament formed with ATP, which is presumably because the filament disassembles (Fig. 1c, lanes 7 and 8). Synthesis by pol III is more processive on the undamaged oligomer template in the presence of β,γ -complex plus SSB (Fig. 1c, lanes 4 and 8) (see also ref. 9), but it fails to catalyse TLS in either the presence or absence of RecA and/or SSB (data not shown). Thus, the absolute requirement for SSB in copying the stabilized RecA filament involves specific interactions between SSB, RecA and pol V. TLS is abolished in the presence of RecA1730 (ref. 1), a missense mutant refractory to Umu-dependent mutagenesis¹⁰, which shows a direct function for RecA in this process.

TLS and primer-use efficiencies were measured at different concentrations of each of the three pol V Mut accessory factors in the presence of fixed levels of the remaining two (Fig. 2; see also Supplementary Information). A rapid increase in TLS and primer use occurs as the concentration of RecA is increased from 0 to about 300 nM, followed by a levelling off in TLS and a shallow decline in the fraction of extended primers (Fig. 2a). The two peaks at 300 nM RecA occur at a RecA/single-stranded DNA (ssDNA) ratio of about 1 RecA per 3–4 nucleotides, which is roughly consistent with the optimal ratio for RecA filament formation⁸. As in the case of RecA, TLS increases sharply with increasing SSB levels, reaching a peak at about 60 nM SSB, followed by a levelling off in the lesion-bypass efficiency (Fig. 2b). The increase in primer use is much less steep, and peaks at roughly 400 nM SSB.

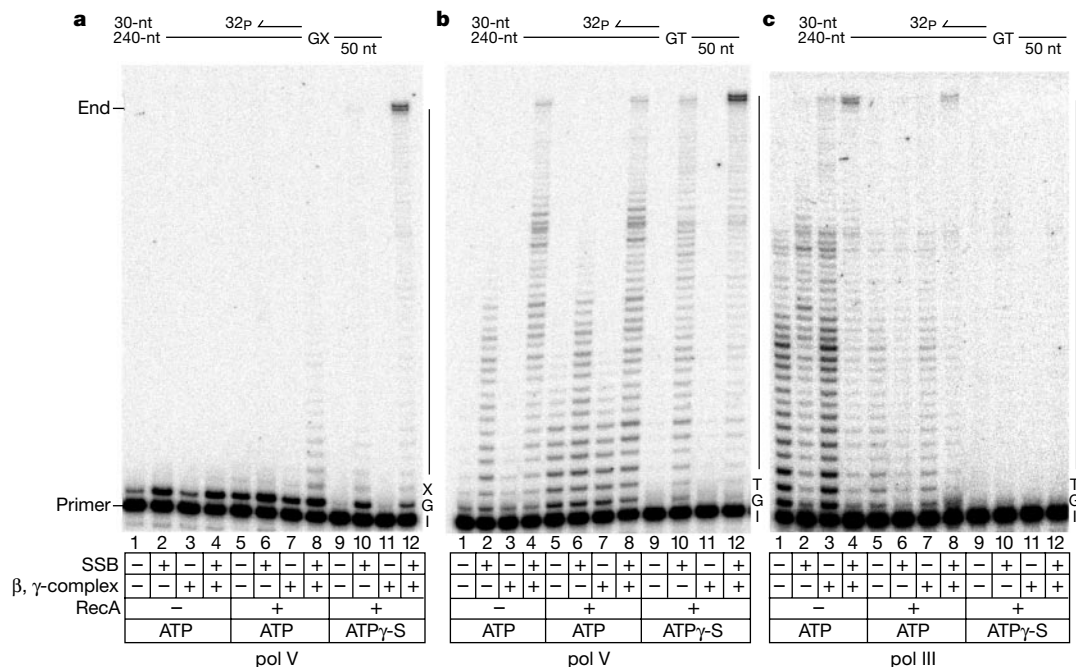


Figure 1 Effect of RecA, SSB and β,γ -complex on replication of lesion-containing and natural primer/template (p/t) DNA. Pol V catalysed DNA synthesis on lesion-containing p/t DNA (**a**) and on natural p/t DNA (**b**). **c**, Pol III catalysed DNA synthesis on natural p/t DNA.

The p/t DNA consists of a ³²P-labelled 30-nucleotide (nt) primer annealed to a 240-nt DNA template containing either an abasic lesion (X) or a natural base (T) (top).

We performed steady-state kinetics experiments to ascertain the individual effects of each of the three mutasomal components (RecA, SSB and β,γ -complex) on pol V activity (Table 1). ATP was used in the reactions because in the absence of SSB negligible amounts of synthesis occurred with ATP- γ S (Fig. 1a, b, lane 11). The addition of RecA and SSB proteins strongly stimulated pol V activity by 340- and 1,040-fold, respectively, whereas β,γ -complex exerted threefold stimulation (Table 1). The stimulatory effects of RecA and β,γ -complex are due to a reduction in the apparent $K_{m,dNTP}$ values, with no measurable change in the apparent V_{max} , whereas there is a 150-fold decrease in $K_{m,dNTP}$ and a sevenfold increase in V_{max} accompanying stimulation by SSB (Table 1). (The apparent Michaelis constant $K_{m,dNTP}$ is defined as the dNTP substrate concentration required to attain one-half the maximum nucleotide incorporation

velocity, V_{max} , at a specific concentration of primer/template DNA.) These data indicate that stimulation in pol V synthesis in the presence of RecA, SSB and β,γ -complex may be due to an increased affinity for template-primer 3' ends. The synthesis pattern taking place on a stabilized RecA filament (using ATP- γ S) suggests that β,γ -complex increases pol V processivity (Fig. 1a and b, compare lanes 10 and 12). A negligible amount of TLS occurs in the absence of β -sliding clamp (Figs 1a, lane 10 and 2c); however, increased processivity ensues when β,γ -complex levels are increased, with TLS efficiency achieving a maximum when β is present at roughly 400 nM.

The increase in processive synthesis observed using RecA filaments formed with ATP- γ S provides an insight into the mechanism of lesion bypass catalysed by pol V Mut. The diameter of the RecA

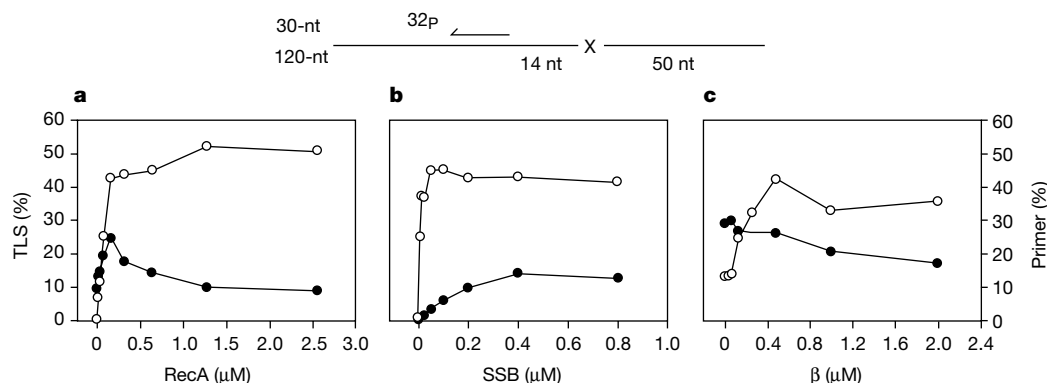


Figure 2 Dependence of pol V-catalysed TLS and total DNA synthesis on RecA, SSB and β,γ -complex. **a**, RecA protein concentration was varied while maintaining constant concentrations of SSB (400 nM), β - (400 nM) and γ -complex (100 nM). **b**, SSB protein concentration was varied while maintaining constant concentrations of RecA (250 nM) and β - (400 nM), γ -complex (100 nM). **c**, The concentration of the β -dimer was varied

while coordinately varying the concentration of the γ -clamp loading complex between 0 and 0.5 μ M (at a 4/1 β/γ ratio), while maintaining constant concentrations of RecA (250 nM) and SSB (400 nM). ATP- γ S (1 mM) was present in all reactions. Open circles represent the efficiency of abasic TLS; closed circles represent the efficiency of primer extension (primer use). The abasic moiety is indicated (X).

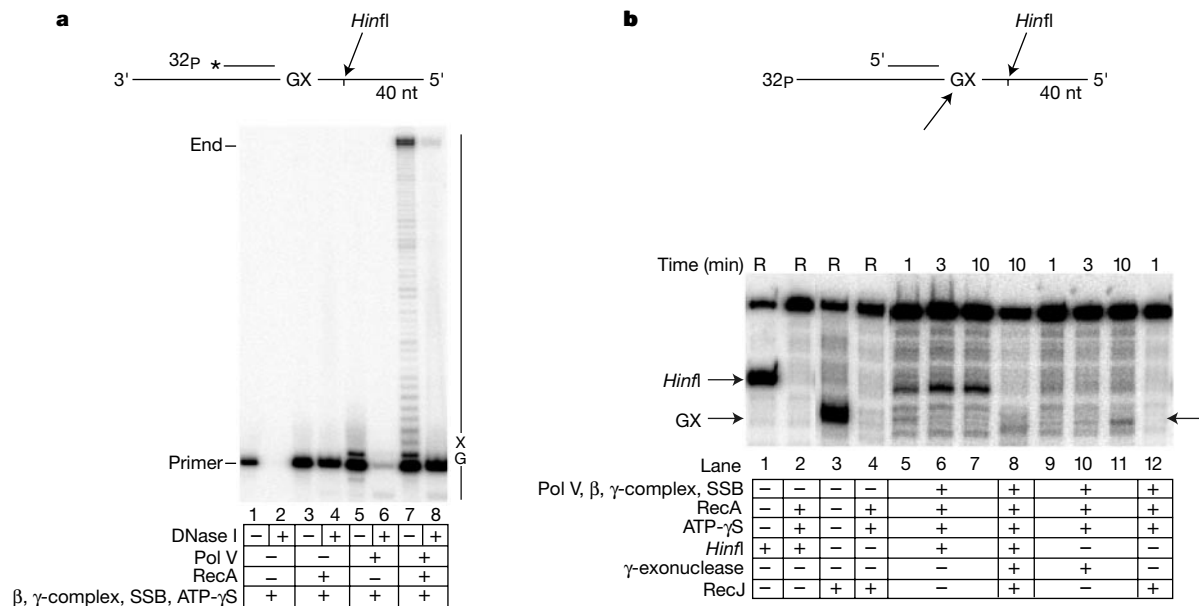


Figure 3 Nuclease protection analysis of pol V plus SSB 3' to 5' disassembly of a stabilized RecA filament. A RecA filament assembled with ATP- γ S is copied using pol V Mut and then digested with various combinations of nucleases. **a**, DNase I degrades a 32 P 5'-end-labelled primer of 30 nt annealed to an unlabelled 240-nt template. Synthesis with pol V Mut was carried out for 10 min (lanes 5–8). **b**, Combinations of HinfI restriction endonuclease, RecJ single-stranded 5' to 3' exonuclease and λ -double-stranded 5' to 3' exonuclease used to digest a 32 P 3'-end-labelled template of 240 nt annealed to an

unlabelled 30-nt primer. Synthesis by pol V Mut was carried out for 1, 3 and 10 min, resulting in incorporation of C opposite G to reach the template abasic lesion (X), followed by the addition of 10 nt downstream from X to copy past the HinfI restriction site, and continuing for an additional 40 nt to reach the 5' end of the template. Four control reactions (R) are shown. For the HinfI control (lanes 1 and 2), the 30-nt primer is replaced with a 63-nt primer covering the HinfI restriction site. A sketch of the p/t DNA is shown (top).

filament is roughly 100 Å (ref. 8), whereas the inner diameter of the doughnut-shaped β -dimer sliding clamp is only 35 Å (ref. 11). Thus, it seems unlikely that processive synthesis could occur by threading the filament through the much narrower opening in the β -clamp. We show, instead, that RecA molecules are 'stripped' from the filament by the combined action of pol V plus SSB. The stripping reaction releases RecA in a 3' to 5' direction (with respect to the template strand) immediately ahead of the advancing polymerase.

Notably, pol V plus SSB drives RecA filament disassembly in a 3' to 5' direction. The process of filament stripping can be dissected using a combination of endo- and exonucleases to digest elongated primer and replicated template strands (Fig. 3). We first examined the ability of DNase I to degrade a ^{32}P 5'-end-labelled primer annealed to an unlabelled template (Fig. 3a), and then we used RecJ and λ -exonucleases in conjunction with *HinfI* restriction nuclease to digest a ^{32}P 3'-end-labelled template annealed to an unlabelled primer (Fig. 3b).

We carried out a DNase I protection assay on replicated and non-replicated stabilized RecA template filaments (Fig. 3a, lanes 7 and 3, respectively). Synthesis by pol V proceeds to the end of the RecA filament (Fig. 3a, lane 7). When a RecA filament assembles on primer/template DNA that contains a 5'-single-strand template overhang (Fig. 3), it continues growing in a 5' to 3' direction, coating the double-stranded DNA (dsDNA) region where the primer and template have annealed¹². The RecA-coated primer is therefore protected (that is, it is resistant to digestion by DNase I) in the absence of pol V (Fig. 3a, lane 4), whereas the naked DNA is destroyed (Fig. 3, lane 2). In contrast, > 92% of the pol V extension products that have been ^{32}P -labelled are degraded (Fig. 3a, lane 8), as are almost all (> 98%) of the non-extended primer strands in the absence of RecA (Fig. 3a, lane 6). We conclude, therefore, that RecA has been removed from the template as a consequence of pol V synthesis, requiring the presence of SSB.

The reciprocal experiment, which examines degradation of a ^{32}P 3'-end-labelled template and unlabelled primer, shows that RecA is stripped in the 3' to 5' direction (Fig. 3b). The appearance of a *HinfI* restriction band after a 1-min reaction with pol V Mut (Fig. 3b, lane 5) shows that at least 16 nucleotides are incorporated when copying the abasic lesion (X) and the downstream restriction site. The control shows that *HinfI* fails to cut a RecA-coated primer/template DNA that has a double-stranded region extending 20 nucleotides beyond the restriction site (Fig. 3b, lane 2). However, *HinfI* is able to restrict the control DNA in the absence of RecA (Fig. 3b, lane 1). The absence of a *HinfI* restriction band in the control (Fig. 3b, lane 2) confirms that a RecA filament assembled on the 5'-template overhang also coats the double-stranded region. Thus, the ability of *HinfI* to cut the newly replicated DNA (Fig. 3b, lane 5) confirms the absence of RecA on the replicated portion of the template, and is consistent with DNase I digestion of the primer (Fig. 3a).

An increase in the restriction-band intensity at 3 min reflects the increased number of primers that extend beyond the *HinfI* site

(Fig. 3b, lane 6). No further increase is observed with a longer, 10-min, incubation (Fig. 3b, lane 7). The restriction band disappears after incubation of the *HinfI* digest with both RecJ, which digests ssDNA 5' to 3', and λ -exonuclease, which digests dsDNA 5' to 3' (Fig. 3b, lane 8). We conclude that RecA has been stripped from the 3'-portion of the template strand beginning at G and continuing beyond the *HinfI* site (Fig. 3b).

Although a 1-min reaction with pol V Mut is sufficient to copy past the restriction site, synthesis does not reach the end of the template. The failure of λ -exonuclease to digest the 5' end of the template at the 1-min time point (Fig. 3b, lane 9) suggests that the DNA is not yet double stranded; however, digestion with λ -exonuclease does take place (Fig. 3b, lanes 10 and 11) once synthesis has reached the end of the template at 3 and 10 min (for example see Fig. 3a, lane 7). An increase in the intensity of the λ -exonuclease digestion band at 10 min compared with at 3 min indicates an increased number of primers reaching the end of the template strand, which further substantiates the hypothesis that RecA is stripped in a 3' to 5' direction ahead of pol V. The template strand is refractory to digestion with RecJ exonuclease at the 1-min time point (Fig. 3b, lane 12), which shows that the single-stranded 5'-portion of the template RecA filament has remained intact. Thus, stripping of RecA does not occur in a 5' to 3' direction. As expected, RecJ does not degrade the template strand at either 3 or 10 min because it is inactive on RecA-coated ssDNA and on naked dsDNA (data not shown).

Our proposed model for TLS catalysed by pol V Mut addresses events that occur subsequent to the binding of pol V at a primer-end proximal to a template lesion (Fig. 4). The unique feature of the model is that pol V, acting together with SSB, behaves in a manner that is loosely analogous to a 'cowcatcher' (a V-shaped device attached to the front of a locomotive to push obstacles off the track). The observation that synthesis by the pol V plus SSB locomotive is processive on a stabilized RecA filament, but only when β , γ -complex is present (Fig. 1a and b, lane 12), provides an important mechanistic insight: a 100-Å RecA filament cannot thread through a 35-Å diameter β -sliding-clamp 'doughnut' hole. Therefore, pol V plus SSB must first release the RecA from the template strand (Fig. 4), in a reaction that does not require ATP hydrolysis. A probable function of β , γ -complex during TLS may be to provide additional stability to the pol V/DNA complex, thereby helping pol V to maintain contact with the tip of the receding filament¹³. However, the cowcatcher analogy should not be taken literally. In contrast to displaced 'living' cows, RecA is not present as a passive obstacle to the advancing pol V. Rather, RecA is an active participant during TLS—it is responsible for stimulating pol V activity (Table 1), and in its absence TLS cannot occur (Fig. 1a).

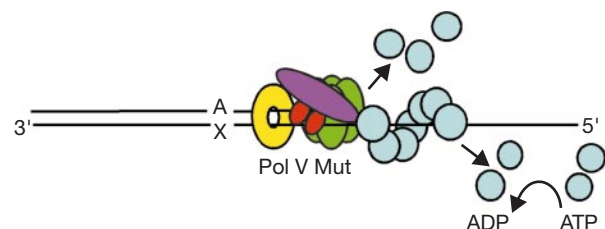


Figure 4 Cowcatcher model for TLS by the pol V mutasome. TLS catalysed by the pol V mutasome is indicated by the insertion of A opposite X. Pol V, in conjunction with SSB, catalyses filament disassembly in the 3' to 5' direction by stripping off RecA molecules in a reaction that does not require ATP hydrolysis, while concomitant ATP-hydrolysis-driven filament disassembly occurs in the 5' to 3' direction, resulting in bidirectional filament disassembly. The 5' to 3' filament disassembly reaction causes pol V to lose contact with RecA, helping to confine mutations to the template-damage sites. Pol V (Umu C) is shown in purple (Umu C) and red (Umu D'), Rec A in blue, SSB in green and β sliding clamp in yellow.

Table 1 Pol V-catalysed synthesis on an undamaged 120-nt template in combination with RecA, SSB and β , γ -complex*

	$V_{\max}$ (min^{-1})	K_m (μM)	$V_{\max}/K_m$ ($\mu\text{M}^{-1} \text{min}^{-1}$)	Fold stimulation*
Pol V	0.2	411	0.0005	1
Pol V + RecA	0.2	1.2	0.17	340
Pol V + SSB	1.4	2.7	0.52	1,040
Pol V + β , γ	0.2	157	0.0013	3
Pol V + RecA + SSB + β , γ -complex	0.5	0.9	0.56	1,220

The kinetic data refer to the incorporation of dAMP opposite T using the 30/120-nt p/t DNA. p/t DNA (2.5 nM) was incubated with RecA (64 nM), β - (40 nM), γ -complex (10 nM), SSB (250 nM) and ATP (1 mM) at 37 °C for 5 min. The reactions were initiated by addition of Pol V (240 nM) and various dATP substrate concentrations. Reactions were quenched after 6 min. Apparent $V_{\max}$ and K_m values were calculated as described¹⁹.

* Fold stimulation is calculated by dividing $V_{\max}/K_m$ for Pol V obtained with various combinations of mutasomal proteins by the value obtained with Pol V alone.

The mechanisms responsible for targeting pol V to the site of template damage are unclear. Unlike nucleotide excision–repair systems that depend on proteins to scan the DNA in search of lesions¹⁴, TLS has a ‘built-in’ lesion-targeting mechanism, a stalled replication fork proximal to a damaged template base¹⁵. Even so, SSB, and perhaps even pol III, may have a hand in escorting pol V to the lesion. SSB has a higher affinity (threefold) for binding to ultraviolet-damaged DNA¹⁶, so that targeting may occur by means of a pol V/SSB interaction. Interactions between replicative pol III holoenzyme and pol V have been observed^{2,17}, which raises the possibility that pol V and pol III may both be part of a higher order replisome that is designed to deal effectively with a variety of DNA damage. □

Methods

Materials

We purchased ATP- γ S from Roche, and DNase I (7,500 units ml⁻¹) was from Amersham-Pharmacia. We purchased *Hinf*I restriction endonuclease (10,000 units ml⁻¹) and λ -exonuclease (5,000 units ml⁻¹) from New England Biolabs. We purified RecJ exonuclease essentially as described¹⁸. Proteins and other chemicals are described in ref. 4. Templates used in this study were 240- and 120-nucleotide synthetic oligonucleotides with or without a synthetic abasic (tetrahydrofuran) lesion. We used ³²P 5'-end-labelled or unlabelled 30-nucleotide sequences as primers.

Nucleotide incorporation on DNA primer templates

All reaction mixtures (10 μ l) contained 20 mM Tris (pH 7.5), 8 mM MgCl₂, 5 mM dithiothreitol, 0.1 mM EDTA, 25 mM sodium glutamate, 40 μ g ml⁻¹ bovine serum albumin and 4% (v/v) glycerol. Reactions were carried out at 37 °C in the presence of primer/template DNA (10 nM), pol V (480 nM) or pol III core (5 nM), ATP (1 mM) or ATP- γ S (1 mM) and the four dNTP substrates (200 μ M each).

RecA protein (750 nM), SSB (500 nM), β - (160 nM) and γ -complex (40 nM) were permuted as indicated in Fig. 1. The reactions were carried out at 37 °C for 10 min and terminated by adding 20 μ l of 20 mM EDTA in 95% formamide. Synthesis products were heat denatured, separated on a 12% polyacrylamide denaturing gel, and band intensities were measured by phosphorimaging using ImageQuant software (Molecular Dynamics). Primer/template DNA was computed from the integrated gel-band intensities as the fraction of extended primers in the reaction. TLS values were calculated as the number of primers extended directly opposite the lesion plus the number of primers extended past the lesion, divided by the total number of primers extended. Nucleotide-incorporation kinetics measurements were from standing-start reactions as described¹⁹.

Nuclease protection analyses

In the DNase I protection assay, primer/template DNA (10 nM) was pre-incubated in either the presence or absence of RecA (750 nM) along with SSB (500 nM), β - (400 nM), γ -complex (100 nM), pol V (480 nM) and ATP- γ S (1 mM) for 5 min at 37 °C. We added the four dNTP substrates (200 μ M each) to initiate the reaction, which we carried out for 10 min. DNase I digestion was performed by mixing an aliquot (5 μ l) from each reaction with DNase I (30 units). After a 1-min incubation, the reactions were quenched by addition of 95% formamide (20 μ l) and EDTA (20 mM), and heated at 95 °C for 10 min. Replicated and non-replicated 30/240-nucleotide primer/template DNAs, in the presence or absence of RecA, were digested with *Hinf*I restriction endonuclease in combination with RecJ (5' to 3') single-stranded exonuclease and λ -exonuclease (5' to 3') double-stranded exonuclease, using reaction and quenching conditions identical to the DNase I protection assay. After adding the four dNTP substrates to initiate the reaction, aliquots (5 μ l) were removed after 1, 3 and 10 min, and transferred to reaction containing *Hinf*I (1 unit), λ -exonuclease (1 unit) and RecJ exonuclease (0.6 ng) either alone or in combination, as indicated in Fig. 3b.

Received 10 October; accepted 23 November 2000.

1. Tang, M. *et al.* Biochemical basis of SOS mutagenesis in *Escherichia coli*: reconstitution of *in vitro* lesion bypass dependent on the UmuD₂C mutagenic complex and RecA protein. *Proc. Natl Acad. Sci. USA* **95**, 9755–9760 (1998).
2. Tang, M. *et al.* UmuD₂C is an error-prone DNA polymerase, *Escherichia coli* pol V. *Proc. Natl Acad. Sci. USA* **96**, 8919–8924 (1999).
3. Reuven, N. B., Arad, G., Maor-Shoshani, A. & Livneh, Z. The mutagenic protein UmuC is a DNA polymerase activated by UmuD₂, RecA, and SSB and is specialized for translesion replication. *J. Biol. Chem.* **274**, 31763–31766 (1999).
4. Tang, M. *et al.* Roles of *E. coli* DNA polymerases IV and V in lesion-targeted and untargeted mutagenesis. *Nature* **404**, 1014–1018 (2000).
5. Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* 1 407–522 (ASM, Washington, 1995).
6. Goodman, M. F. & Tiffin, B. The expanding polymerase universe. *Nature Rev. Mol. Cell Biol.* **1**, 101–109 (2000).
7. Echols, H. & Goodman, M. F. Mutation induced by DNA damage: A many protein affair. *Mutat. Res.* **236**, 301–311 (1990).
8. Kuzminov, A. Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage λ . *Microbiol. Mol. Biol. Rev.* **63**, 751–813 (1999).

9. Bloom, L. B. *et al.* Dynamics of loading the β sliding clamp of DNA polymerase III onto DNA. *J. Biol. Chem.* **271**, 30699–30708 (1996).
10. Bailone, A., Sommer, S., Knezevic, J., Dutreix, M. & Devoret, R. A RecA protein mutant deficient in its interaction with the UmuDC complex. *Biochimie* **73**, 479–484 (1991).
11. Kong, X.-P., Onrust, R., O'Donnell, M. & Kuriyan, J. Three-dimensional structure of the β subunit of *E. coli* DNA polymerase III holoenzyme: a sliding DNA clamp. *Cell* **69**, 425–437 (1992).
12. Shan, Q., Bork, J. M., Webb, B. L., Inman, R. B. & Cox, M. M. RecA protein filaments: end-dependent dissociation from ssDNA and stabilization by RecO and RecR proteins. *J. Mol. Biol.* **265**, 519–540 (1997).
13. Sommer, S., Bailone, A. & Devoret, R. The appearance of the UmuD₂C protein complex in *Escherichia coli* switches repair from homologous recombination to SOS mutagenesis. *Mol. Microbiol.* **10**, 963–971 (1993).
14. Lindahl, T. & Wood, R. D. Quality control by DNA repair. *Science* **286**, 1897–1905 (1999).
15. Goodman, M. F. Coping with replication ‘train wrecks’ in *Escherichia coli* using pol V, pol II, and RecA proteins. *Trends Biochem. Sci.* **25**, 189–195 (2000).
16. Lao, Y., Gomes, X. V., Ren, Y., Taylor, J. S. & Wold, M. S. Replication protein A interactions with DNA. III. Molecular basis of recognition of damaged DNA. *Biochemistry* **39**, 850–859 (2000).
17. Sutton, M. D., Opperman, T. & Walker, G. C. The *Escherichia coli* SOS mutagenesis proteins UmuD and UmuD₂ interact physically with the replicative DNA polymerase. *Proc. Natl Acad. Sci. USA* **96**, 12373–12378 (1999).
18. Lovett, S. T. & Kolodner, R. D. Identification and purification of a single-stranded-DNA-specific exonuclease encoded by RecJ gene of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **86**, 2627–2631 (1989).
19. Creighton, S., Bloom, L. B. & Goodman, M. F. In *Methods Enzymol.* (ed. Campbell, J. L.) 232–256 (Academic, San Diego, 1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This work was supported by National Institutes of Health grants to M.F.G. and M.O. P.P. was supported on an NIH-NIA postdoctoral training grant, and J.G.B. was supported on a National Institute of Dental and Craniofacial Research predoctoral training grant.

Correspondence and requests for materials should be addressed to M.F.G. (e-mail: mgoodman@mizar.usc.edu).

χ-Sequence recognition and DNA translocation by single RecBCD helicase/nuclease molecules

Kathleen M. Dohoney & Jeff Gelles

Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02454-9110, USA

Major pathways of recombinational DNA repair in *Escherichia coli* require the RecBCD protein—a heterotrimeric, ATP-driven, DNA translocating motor enzyme. RecBCD combines a highly processive and exceptionally fast helicase (DNA-unwinding) activity with a strand-specific nuclease (DNA-cleaving) activity (refs 1, 2 and references therein). Recognition of the DNA sequence ‘ χ ’ (5'-GCTGGTGG-3') switches the polarity of DNA cleavage and stimulates recombination at nearby sequences *in vivo*. Here we attach microscopic polystyrene beads to biotin-tagged RecD protein subunits and use tethered-particle light microscopy to observe translocation of single RecBCD molecules (with a precision of up to ~30 nm at 2 Hz) and to examine the mechanism by which χ modifies enzyme activity. Observed translocation is unidirectional, with each molecule moving at a constant velocity corresponding to the population-average DNA unwinding rate. These observations place strong constraints on possible movement mechanisms. Bead release at χ is negligible, showing that the activity modification at χ does not require ejection of the RecD subunit from the enzyme as previously proposed; modification may occur through an unusual, pure conformational switch mechanism.

Conventional techniques for studying enzyme function are restricted to observing the population-averaged properties of molecular ensembles. By contrast, light-microscopy methods can follow

translocating single enzyme molecules along DNA with measurement precision of the order of ~ 10 nm (ref. 3 and references therein). To examine translocation on DNA by single RecBCD molecules, we first prepared a derivative of RecBCD, RecBCD-bio, which has a biotin moiety near the carboxy terminus of the RecD subunit. RecBCD-bio expression rescues the ultraviolet sensitivity⁴ of *recD⁻recJ⁻* cells (data not shown), showing that the enzyme is functional *in vivo*. Purified RecBCD-bio exhibits nuclease, helicase and DNA-dependent ATPase activities *in vitro* analogous to those of the wild-type enzyme.

Single RecBCD-bio molecules were labelled with streptavidin-coated, 199-nm diameter polystyrene beads (see Methods) for use in a light-microscopy tethered particle motion (TPM) experiment^{5,6}, in which binding to and translocation along DNA by the enzyme are observed by monitoring changes in bead brownian motion (Fig. 1a; see Supplementary Information). When introduced above a glass surface with attached DNA molecules in the presence of 50 μ M ATP, individual enzyme-bead complexes exhibit rapid brownian motion and typically diffuse out of the microscope focal plane in less than 1 s. Occasionally, the complexes suddenly adopt the restricted brownian motion characteristic of DNA-tethered particles (Fig. 1b, c)⁵⁻⁷. The range of brownian motion then progressively decreases over about 10 s until visible brownian motion ceases. Most beads subsequently abruptly detach from the surface and diffuse out of the microscope focal plane (Fig. 1c-f). This behaviour is that expected for enzyme-bead complexes that bind to the free ends of surface-bound DNA molecules, translocate along and unwind the DNA, and then detach from the single-stranded products when the

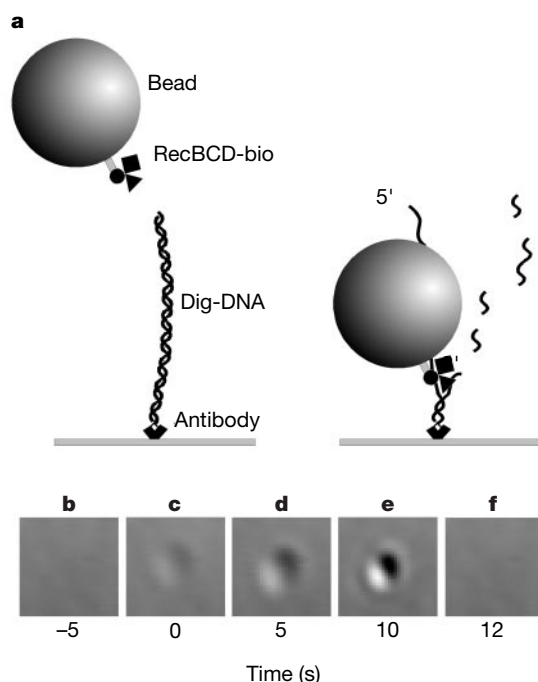


Figure 1 Observing translocation along DNA by single RecBCD-bio molecules.

a, Experimental design (not to scale). Duplex DNA molecules (1.3–1.5 kb), labelled with digoxigenin at one end (dig-DNA), attach to a glass surface coated with anti-digoxigenin IgG (antibody). RecBCD-bio molecules labelled with microscopic polystyrene beads (left) bind to and begin unwinding the free DNA ends. The enzyme may cleave one or both of the resulting single-stranded ends as unwinding progresses (right). The duplex DNA acts as a flexible tether; as it shortens there is an observable decrease in the spatial extent of bead Brownian motion. **b–f**, Time-averaged (over 0.5 s) video images of a $2.4 \times 2.4 \mu$ m area of the microscope field. At time zero, a bead became tethered; subsequent decrease in Brownian motion owing to tether shortening is seen as a gradual decrease in blur and increase in contrast in the images (**c–e**). At time 11.5 s, the bead detached from the surface and diffused away.

opposite end of the DNA is reached.

A progressive decrease in brownian motion was not observed in the absence of enzyme, or when enzyme activity was blocked by withholding ATP. About 30% of the tethered beads do not translocate; a minority fraction of non-productive complexes is also detected in macroscopic pre-steady-state kinetics experiments

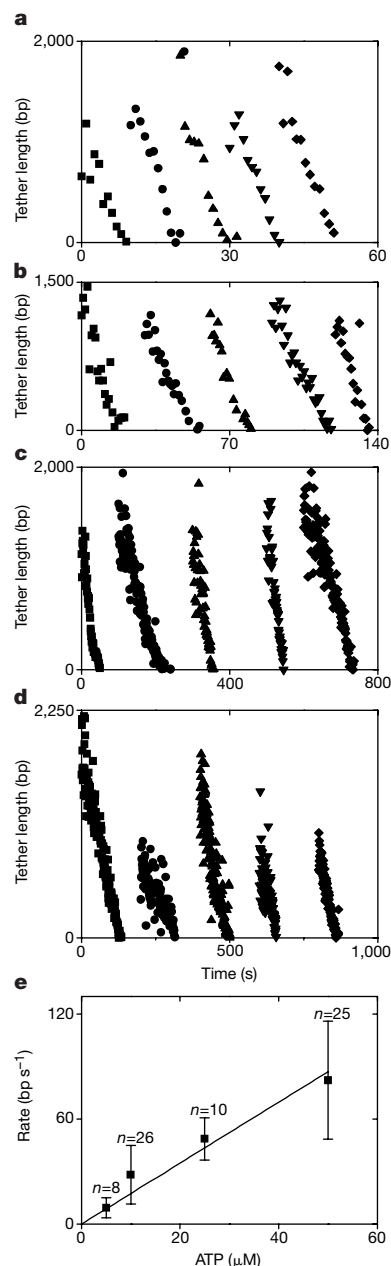


Figure 2 Time courses and rates of translocation. **a–d**, Examples of single-molecule RecBCD-bio translocation time courses at 22–25 °C: 50 μ M (**a**), 25 μ M (**b**), 10 μ M (**c**) or 5 μ M (**d**) ATP; note the different timescales. Each panel shows records of five different single enzyme molecules translocating on DNA molecules (~ 1.3 kb), which contained χ sequences only in the reverse orientation (which is not recognized by RecBCD) or lacked χ altogether. Measurement precision increases as tether length decreases⁶; precision is 114 bp (40 nm) r.m.s. at a tether length of 208 bp. Initial tether lengths vary owing to systematic errors in measurement⁶ and to small differences in the lengths of the various DNA constructs used. Time zero for each record is arbitrary. **e**, Population mean translocation rates ($\pm$ s.d.) calculated from n single-molecule tether-length time courses. All ATP concentrations are $\ll K_M$ for DNA unwinding (130 μ M)⁹; thus, data are fit to a proportional model (line) with fit slope corresponding to the translocation $V_{max}/K_M = 1.7 \pm 0.1$ bp s⁻¹ μ M⁻¹.

(A. Lucius and T. Lohman, personal communication). The observed surface densities of tethered beads were more than 25-fold lower in control experiments in which the antibody-coated surface was pre-treated with excess digoxigenin-UTP before the addition of DNA (1.1×10^{-3} versus $< 3.5 \times 10^{-5}$ tethered beads per μm^2), or the streptavidin-coated beads were pre-treated with excess biotin before the addition of RecBCD-bio (4.6×10^{-3} versus 1.8×10^{-4} tethered beads per μm^2 ; $> 0.028 \text{ mm}^2$ examined in each sample). These experiments show that bead tethering arises predominantly from specific attachment of the biotinylated RecD subunit and DNA to the beads and glass surface, respectively.

Using image-processing techniques^{6,7}, we measured the position time courses of 69 individual enzyme molecules moving on DNA over a range of subsaturating ATP concentrations (Fig. 2). Most enzyme molecules translocate at a constant rate as they travel along the DNA, which is more than 1 kb in length. A minority exhibit a slight, gradual decrease in velocity at the end of the record (for example, Fig. 2c, 2nd trace), perhaps owing to unfavourable mechanical interactions between the bead and the chamber surface when the tether is very short. Sixty-eight of the sixty-nine molecules showed no detectable lag (detection limit: 2.5–20 s, depending on translocation velocity) between DNA binding and commencement of translocation. Thus, RecBCD lacks the substantial kinetic barrier associated with duplex opening during initiation, which is seen, for example, with RNA polymerases.

Translocation is continuous to the limit of experimental detection; transient cessation of movement was observed in only one data record. The velocity distribution at each ATP concentration was unimodal and roughly gaussian in shape, with standard deviation (error bars in Fig. 2e) about that expected from the uncertainty in position measurements alone⁶. Thus, to the limits of experimental resolution, the translocating enzyme molecules are a uniform, kinetically homogeneous population. Observed single-molecule translocation rates (Fig. 2e) are identical to ensemble-average RecBCD DNA unwinding rates ($V_{\text{max}}/K_M = 1.92 \pm 0.6 \text{ bp s}^{-1} \mu\text{M}^{-1}$) measured under similar conditions^{8,9}. Thus, translocation and unwinding are

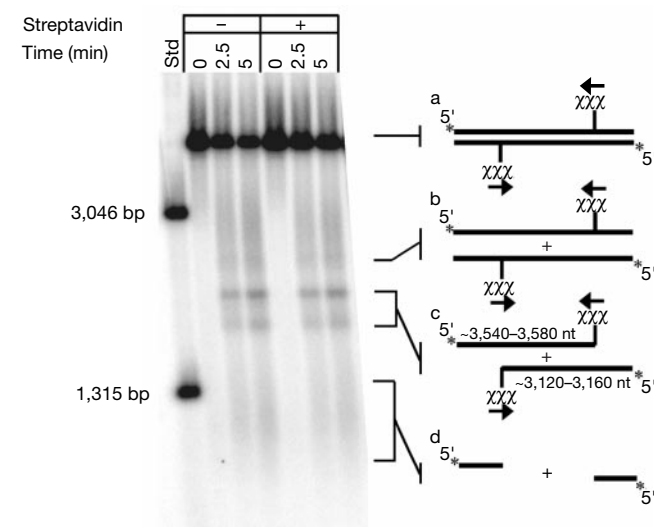


Figure 3 Non-denaturing electrophoresis gel autoradiogram showing production of χ -specific DNA fragments by RecBCD-bio. Where indicated, enzyme was pre-incubated with 100-fold molar excess streptavidin. Std, double-stranded molecular weight markers. RecBCD-bio is reacted (see Methods) for varying times with ^{32}P 5'-end-labelled (asterisk) substrate DNA (a); substrate is unwound and simultaneously partially degraded until nuclease activity is altered on recognition of a correctly orientated χ sequence, producing χ -specific products (c). Shorter products (d) are the result of a χ -specific nick in the DNA strand opposite the recognized χ ³⁰. Undegraded full-length single-stranded products (b) are produced in a small fraction of the unwinding events.

tightly coupled over a wide range of ATP concentrations.

Processive motor enzymes differ in the extent to which the translocation mechanism requires large-scale random motion of the motor on its track. In one type of mechanism, exemplified by that of the microtubule motor KIF1A¹⁰, the enzyme slides randomly forward and backward over distances more than 1 μm ; net forward movement occurs through a small bias in this random walk. The opposite extreme is an enzyme like conventional kinesin¹¹, another microtubule motor, which moves unidirectionally in a series of discrete, nanometre-scale jumps or 'steps' of uniform size. In our experiments, there is no evidence for significant backward translocation of RecBCD. In the short-tether portions of the 5- μM ATP records (these contain the most precise tether-length measurements⁶ of the most slowly translocating molecules—the conditions in which backward movements, if present, would be most readily detected), 0 of 203 movements over $\sim 0.5 \text{ s}$ were backward by a distance exceeding two standard deviations ($\sim 100 \text{ nm}$) of the measurement precision⁶, as compared with 5 that would be expected from gaussian noise alone. Thus, RecBCD translocation resembles that of kinesin more closely than that of KIF1A; if there is any random-walk component of RecBCD movement, it must occur only over very short distances and/or be highly

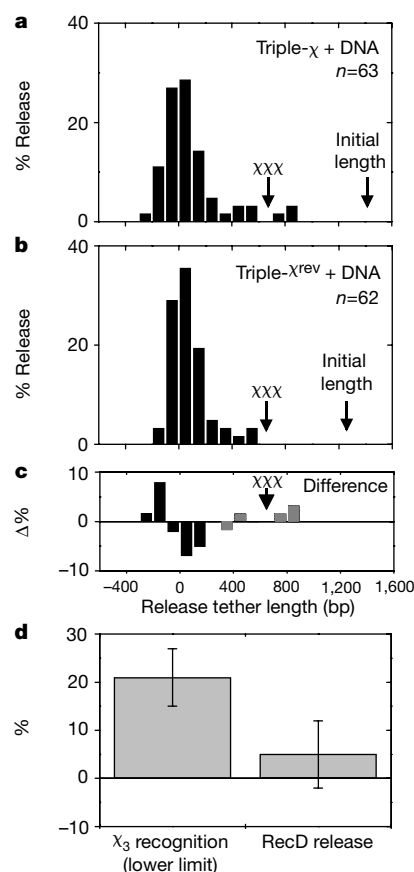


Figure 4 RecD release at χ . **a–c**, Release of RecD subunits at various positions on DNA in single-molecule translocation experiments. DNAs contain a triple tandem repeat of the χ orientated correctly (a) or incorrectly (b) for recognition. Arrows denote the midpoint of the triple χ repeat at 638 bp and the initial DNA tether length of each substrate. The difference (c) between the fraction of release events on the forward- and reverse-orientated templates represents the events that are specifically due to χ recognition; $4 \pm 7\%$ (mean $\pm$ s.e.) of the specific events occur within ± 1 tether length measurement standard deviation⁶ of the χ repeat (shaded bars). Measured tether lengths for some release events are less than zero due to experimental uncertainty. **d**, Comparison of χ -specific RecD release measured in single-molecule experiments (c) and the lower limit of χ_3 recognition efficiency measured in ensemble-average experiments (for example, Fig. 3).

biased in the forward direction. These observations argue against RecBCD translocation mechanisms that are analogous to those proposed for RNA polymerase¹², in which the enzyme slides forward and backward along duplex DNA with motion biased in the forward direction by unwinding and/or nuclease reactions that can occur only when the enzyme is positioned adjacent to the single-stranded DNA/double-stranded DNA junction.

A profound, pleiotropic change in enzyme activity occurs in 15–50% of the encounters of RecBCD enzyme with a properly orientated χ sequence (ref. 1 and references therein). Immediately on recognition of the χ , degradation of the 3'-terminal strand is greatly attenuated, 5'-terminal strand degradation is upregulated, and RecA loading activity is stimulated. Unwinding rate and processivity appear unaltered, but ability to initiate the subsequent unwinding of a new duplex can be inhibited. Genetic and biochemical studies have led to the hypothesis that release of the RecD subunit from the holoenzyme is responsible for the χ -induced alteration of activity^{13–17}, but other models have also been proposed^{17–19}.

To measure how efficiently RecBCD-bio recognizes χ , we examined reactions between the enzyme and radiolabelled χ^+ DNA for the presence of χ -specific DNA products²⁰ under experimental conditions (1 mM ATP 1 mM ADP and 1 mM Mg(OAc)₂) optimized for detection of such products at slow²¹ unwinding rates (Fig. 3). The DNA (Fig. 3a) contains triple tandem repeats of χ to increase the fraction of enzyme molecules with χ -modified activity; a χ -specific product of 3.1–3.6 kb (Fig. 3c) will be produced regardless of which end of the DNA the enzyme enters. A χ -specific product was produced from at least $21 \pm 6\%$ (mean $\pm$ s.d., $n = 7$) of the DNA molecules unwound. Binding the enzyme to streptavidin (Fig. 3) or streptavidin-conjugated agarose (data not shown) did not affect significantly the measured recognition efficiency, supporting the hypothesis that the behaviour at χ of RecBCD labelled with streptavidin-conjugated beads is identical to that of the free enzyme. Under conditions (1 mM ATP and 8 mM Mg(OAc)₂) similar to those in previous studies, efficiency of χ recognition by RecBCD-bio (χ_3 recognition efficiency $\geq 28\%$) was similar to that reported^{16,20} for RecBCD.

To test the hypothesis that the modification of RecBCD-bio activity at χ requires loss of the RecD subunit from the holoenzyme–DNA complex, we carried out single-molecule TPM experiments to determine whether release of bead-tagged RecD occurred as the enzyme translocated through correctly orientated χ sequences. Experimental conditions were those used in Fig. 3; the population mean translocation rate was 243 ± 124 bp s⁻¹. No significant pausing or change in velocity at χ was observed. Most beads released from the DNA only on reaching near-zero tether length, indicating that the RecD subunit remained bound to the holoenzyme during translocation through the χ sequences (Fig. 4a). However, a small fraction of the beads ($\sim 13\%$) were released from the DNA as the enzyme traversed a correctly orientated triple χ repeat located near the centre of the DNA. Significantly, a similar, small number of releases near χ was observed when the experiment was repeated using a DNA with reverse-orientated χ sequences, which are not recognized by the enzyme (Fig. 4b)²². The difference in release-position distributions in the vicinity of χ_3 with the forward and reverse orientations gives a measure of χ -site-induced release (Fig. 4c); this quantity is zero within experimental uncertainty (Fig. 4d). That the fraction of enzyme molecules with nuclease activity modified by χ recognition is significantly larger than the small or nonexistent fraction that undergo χ -specific subunit release (Fig. 4d) provides strong evidence that χ modification of enzyme activity does not require ejection of the RecD subunit.

The form of RecBCD with χ -modified nuclease activity is sufficiently kinetically stable that it can persist for hundreds of ATPase catalytic cycles¹⁴. Nevertheless, our data show that RecD dissociation is not required for χ modification of enzymatic activity,

raising the possibility that modification occurs through an unusual pure conformational change mechanism with no linked change in the composition or covalent structure of the enzyme–DNA complex. Alternatively, an unrecognized ligand binding to or post-translational modification of the complex may be responsible for this persistent change in activity. The activity changes studied here, which occur in the RecBCD–DNA complex when the enzyme is positioned at χ , are probably distinct from the changes in the quaternary structure of the enzyme that are detected after RecBCD is no longer bound to double-stranded DNA¹⁷.

The experiments reported here add to existing knowledge of RecBCD function because (1) they directly measure translocation as opposed to unwinding; (2) they do so with a precision approaching ~ 100 bp (ref. 6); (3) they avoid information loss through population averaging by examining the behaviour of isolated enzyme molecules rather than the average behaviour of large molecular ensembles; and (4) they directly monitor the presence of the RecD subunit during DNA translocation. The data show that individual RecBCD-bio molecules move at roughly constant velocity corresponding to the RecBCD DNA unwinding rate over a wide range of ATP concentrations. Movement is mostly unidirectional, supporting the proposal of a directed-stepping model for translocation. Together, our results place significant constraints on the range of possible movement mechanisms. Additional insight into the mechanism could be obtained if future single-molecule studies can improve the measurement precision sufficiently to resolve the unitary steps associated with individual enzymatic cycles. □

Methods

Plasmid to express biotinated RecD

To construct plasmid pPB800-bio, encoding a fusion of RecD protein with the carboxy-terminal 87 amino acids of the *E. coli* biotin carboxyl carrier protein²³, an *Xma*I site was created by directed mutagenesis (5'-CACGGG-3' to 5'-CCCGGG-3') at the C-terminal end of *recD* in pPB800 (a gift from P. Boehrmer)²⁴. The 308-bp polymerase chain reaction (PCR) product amplified from pCY142 (ref. 23) with primers 5'-CGGTACCCGGGATCCTC-3' and 5'-CGATTCCCGGGAATCCATGGAAGCGCCAGCAG-3' was digested with *Xma*I and ligated into the *Xma*I site of the mutated pPB800.

Expression and purification of RecBCD-bio

E. coli JM109 co-transformed with pPB800-bio and pPB520 (a gift from P. Boehrmer)²⁴ was plated on minimal media containing 0.4% glucose, $10 \mu\text{g ml}^{-1}$ thiamine, $100 \mu\text{g ml}^{-1}$ ampicillin and $10 \mu\text{g ml}^{-1}$ chloramphenicol, and grown in 2YT supplemented with 0.1 mM biotin, induced with IPTG, collected as described²⁴ and stored at -20°C . Frozen cells were lysed²⁵ and the lysate dialysed and applied to a monomeric avidin column as described²⁶. The dialysis was in reaction buffer (25 mM Tris-acetate, pH 7.5, 1 mM Mg(OAc)₂, 1 mM dithiothreitol (DTT)) and all column washes were with storage buffer (40 mM Tris-acetate, pH 7.5, 0.2 mM EDTA, pH 7.5, 0.2 mM DTT, 200 mM NaCl). The protein was eluted from the column in the latter buffer supplemented with 5 mM biotin. Peak fractions were pooled, dialysed against storage buffer to remove excess biotin, concentrated in a Centricon-50 (Amicon), diluted 1:2 with glycerol and stored at -20°C . Protein concentration was determined using $\epsilon_{280} = 4 \times 10^5 \text{ M}^{-1} \text{ cm}^{-1}$ (ref. 9). A nitrocellulose blot of an SDS gel of the purified protein showed only one band, the apparent molecular mass of which corresponds to that calculated for RecD-bio, when stained for biotinylated proteins²⁷. Full activation of DNA-stimulated ATPase activity⁹ required 9.8 ± 1.0 heterotrimeric per duplex DNA end, comparable to results obtained with unmodified RecBCD⁹.

Single-molecule translocation measurements

The 1,302–1,478-bp DNA substrates used in the microscope experiments were PCR products of plasmid pBR322 or its χ^+ derivative, pBR322 χ F, 3H (gift of P. Bianco)²⁸, each amplified using one digoxigenin end-labelled primer (Genosys) and one unlabelled primer. Carboxylated polystyrene microspheres (0.199- μm diameter 'beads'; Seradyn) were covalently biotinylated, coated with streptavidin, and purified by gel filtration as described²⁹, except that reaction buffer was used as the gel-filtration buffer. Concentrations of streptavidin-coated bead suspensions were measured as described²⁹. RecBCD-bio was mixed with the streptavidin-coated beads at a ratio that yielded a probability lower than 0.05 of having more than 1 active enzyme per bead, assuming independent binding. Coverslip flow cells (~ 0.1 mm thick) with a volume of $\sim 20 \mu\text{l}$ were prepared as described²⁷. The flow-cell surface was treated with anti-digoxigenin antibody and digoxigenin-labelled DNA as described²⁷, except that washes were $100 \mu\text{l}$ of 1 mg ml⁻¹ filtered casein in reaction buffer and the DNA incubation was 10 min. TPM experiments were initiated by flowing into the cell 50 μl of reaction buffer supplemented with ATP, 16 U ml⁻¹ pyruvate kinase, 1 mM phosphoenolpyruvate, 0.7 μM *E. coli* single-stranded binding protein (SSB), RecBCD-bio–streptavidin bead complex, and 1 mg ml⁻¹ casein

unless otherwise noted. TPM video-enhanced light microscopy, image recording, image processing and data analysis have been described^{5–7}. TPM image acquisition time was 0.5 s.

χ -Recognition efficiency

The DNA substrates were made by treating *Nde*I-linearized pBR322 or pBR322 3 χ F, 3H with calf intestinal phosphatase, followed by 5'-end labelling using T4 polynucleotide kinase and [γ -³²P]ATP. Reactions were conducted at 25 °C essentially as described²⁰ and contained 25 mM Tris-acetate (pH 7.5), 1 mM Mg(OAc)₂, 1 mM DTT, 1 mM ATP, 1 mM ADP, 1.44 nM duplex DNA ends, 1.27 μ M SSB and 0.4 nM active RecBCD-bio. The gels were dried then analysed on a phosphor imager; χ_3 recognition efficiency was taken to be the maximum value over measurements taken at different times of $2P_t/(R_0 - R_t)$ where R_t and P_t are the total radioactivity over background at time t in the unprocessed duplex substrate band and the long χ -specific product bands, respectively. This calculation underestimates the true recognition efficiency because an unknown fraction of the χ -specific products are further degraded during the reaction.

Received 23 August; accepted 7 November 2000.

- Kuzminov, A. Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. *Microbiol. Mol. Biol. Rev.* **63**, 751–813 (1999).
- Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. & Rehauer, W. M. Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol. Rev.* **58**, 401–465 (1994).
- Gelles, J. & Landick, R. RNA polymerase as a molecular motor. *Cell* **93**, 13–16 (1998).
- Lovett, S. T., Luisi-DeLuca, C. & Kolodner, R. D. The genetic dependence of recombination in recD mutants of *Escherichia coli*. *Genetics* **120**, 37–45 (1988).
- Schafer, D. A., Gelles, J., Sheetz, M. P. & Landick, R. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature* **352**, 444–448 (1991).
- Yin, H., Landick, R. & Gelles, J. Tethered particle motion method for studying transcript elongation by a single RNA polymerase molecule. *Biophys. J.* **67**, 2468–2478 (1994).
- Finzi, L. & Gelles, J. Measurement of lactose repressor-mediated loop formation and breakdown in single DNA molecules. *Science* **267**, 378–380 (1995).
- Roman, L. J. & Kowalczykowski, S. C. Characterization of the adenosinetriphosphatase activity of the *Escherichia coli* RecBCD enzyme: relationship of ATP hydrolysis to the unwinding of duplex DNA. *Biochemistry* **28**, 2873–2881 (1989).
- Roman, L. J. & Kowalczykowski, S. C. Characterization of the helicase activity of the *Escherichia coli* RecBCD enzyme using a novel helicase assay. *Biochemistry* **28**, 2863–2873 (1989).
- Okada, Y. & Hirokawa, N. A processive single-headed motor: kinesin superfamily protein KIF1A. *Science* **283**, 1152–1157 (1999).
- Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **365**, 721–727 (1993).
- Guajardo, R. & Sousa, R. A model for the mechanism of polymerase translocation. *J. Mol. Biol.* **265**, 8–19 (1997).
- Thaler, D. S. *et al.* in *Mechanisms and Consequences of DNA Damage Processing* (eds Friedberg, E. & Hanawalt, P.) 413–422 (Alan. R. Liss, New York, 1988).
- Dixon, D. A., Churchill, J. J. & Kowalczykowski, S. C. Reversible inactivation of the *Escherichia coli* RecBCD enzyme by the recombination hotspot chi in vitro: evidence for functional inactivation or loss of the RecD subunit. *Proc. Natl Acad. Sci. USA* **91**, 2980–2984 (1994).
- Myers, R. S., Kuzminov, A. & Stahl, F. W. The recombination hot spot chi activates RecBCD recombination by converting *Escherichia coli* to a recD mutant phenotype. *Proc. Natl Acad. Sci. USA* **92**, 6244–6248 (1995).
- Stahl, F. W., Thomason, L. C., Siddiqi, I. & Stahl, M. M. Further tests of a recombination model in which chi removes the RecD subunit from the RecBCD enzyme of *Escherichia coli*. *Genetics* **126**, 519–533 (1990); erratum *ibid.* **135**, 1232 (1993).
- Taylor, A. F. & Smith, G. R. Regulation of homologous recombination: Chi inactivates RecBCD enzyme by disassembly of the three subunits. *Genes Dev.* **13**, 890–900 (1999).
- Koppen, A., Krobisch, S., Thoms, B. & Wackernagel, W. Interaction with the recombination hot spot chi in vivo converts the RecBCD enzyme of *Escherichia coli* into a chi-independent recombinase by inactivation of the RecD subunit. *Proc. Natl Acad. Sci. USA* **92**, 6249–6253 (1995).
- Anderson, D. G., Churchill, J. J. & Kowalczykowski, S. C. Chi-activated RecBCD enzyme possesses 5' $\rightarrow$ 3' nucleolytic activity, but RecBC enzyme does not: evidence suggesting that the alteration induced by Chi is not simply ejection of the RecD subunit. *Genes Cells* **2**, 117–128 (1997).
- Dixon, D. A. & Kowalczykowski, S. C. The recombination hotspot chi is a regulatory sequence that acts by attenuating the nuclease activity of the *E. coli* RecBCD enzyme. *Cell* **73**, 87–96 (1993).
- Roman, L. J., Eggleston, A. K. & Kowalczykowski, S. C. Processivity of the DNA helicase activity of *Escherichia coli* recBCD enzyme. *J. Biol. Chem.* **267**, 4207–4214 (1992).
- Taylor, A. F., Schultz, D. W., Ponticelli, A. S. & Smith, G. R. RecBC enzyme nicking at Chi sites during DNA unwinding: location and orientation-dependence of the cutting. *Cell* **41**, 153–163 (1985).
- Cronan, J. E. Jr Biotinylation of proteins in vivo. A post-translational modification to label, purify, and study proteins. *J. Biol. Chem.* **265**, 10327–10333 (1990).
- Boehmer, P. E. & Emmerson, P. T. *Escherichia coli* RecBCD enzyme: inducible overproduction and reconstitution of the ATP-dependent deoxyribonuclease from purified subunits. *Gene* **102**, 1–6 (1991).
- Eichler, D. C. & Lehman, I. R. On the role of ATP in phosphodiester bond hydrolysis catalyzed by the RecBC deoxyribonuclease of *Escherichia coli*. *J. Biol. Chem.* **252**, 499–503 (1977).
- Young, E. C., Berliner, E., Mahtani, H. K., Perez-Ramirez, B. & Gelles, J. Subunit interactions in dimeric kinesin heavy chain derivatives that lack the kinesin rod. *J. Biol. Chem.* **270**, 3926–3931 (1995).
- Berliner, E. *et al.* Microtubule movement by a biotinylated kinesin bound to streptavidin-coated surface. *J. Biol. Chem.* **269**, 8610–8615 (1994); erratum *ibid.* **269**, 23382 (1994).
- Anderson, D. G., Churchill, J. J. & Kowalczykowski, S. C. A single mutation, RecB(D1080A) eliminates RecA protein loading but not Chi recognition by RecBCD enzyme. *J. Biol. Chem.* **274**, 27139–27144 (1999).
- Berliner, E., Young, E. C., Anderson, K., Mahtani, H. K. & Gelles, J. Failure of a single-headed kinesin to track parallel to microtubule protofilaments. *Nature* **373**, 718–721 (1995).

- Taylor, A. F. & Smith, G. R. Strand specificity of nicking of DNA at Chi sites by RecBCD enzyme. Modulation by ATP and magnesium levels. *J. Biol. Chem.* **270**, 24459–24467 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>).

Acknowledgements

We thank P. Bianco, P. Boehmer, S. Kowalczykowski, S. Lovett and A. Taylor for materials and helpful advice. This work was supported by the NIGMS.

Correspondence and requests for materials should be addressed to J.G. (e-mail: gelles@brandeis.edu).

Processive translocation and DNA unwinding by individual RecBCD enzyme molecules

Piero R. Bianco^{*†}, Laurence R. Brewer[‡], Michele Corzett[§], Rod Balhorn[§], Yin Yeh^{||}, Stephen C. Kowalczykowski^{*†} & Ronald J. Baskin[†]

Sections of ^{*} Microbiology and of [†] Molecular and Cellular Biology and ^{||} Department of Applied Science, University of California at Davis, Davis, California 95616, USA

[‡] Electronics Engineering Technologies Division, and [§] Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

RecBCD enzyme is a processive DNA helicase¹ and nuclease² that participates in the repair of chromosomal DNA through homologous recombination^{3,4}. We have visualized directly the movement of individual RecBCD enzymes on single molecules of double-stranded DNA (dsDNA). Detection involves the optical trapping of solitary, fluorescently tagged dsDNA molecules that are attached to polystyrene beads, and their visualization by fluorescence microscopy^{5,6}. Both helicase translocation and DNA unwinding are monitored by the displacement of fluorescent dye from the DNA by the enzyme⁷. Here we show that unwinding is both continuous and processive, occurring at a maximum rate of 972 ± 172 base pairs per second ($0.30 \mu\text{m s}^{-1}$), with as many as 42,300 base pairs of dsDNA unwound by a single RecBCD enzyme molecule. The mean behaviour of the individual RecBCD enzyme molecules corresponds to that observed in bulk solution.

Visualization of translocation and DNA unwinding by single DNA helicase molecules permits study of the stochastic properties of individual molecular motors, or 'nano-machines', which are obscured in the population average of steady-state, bulk-phase measurements. To achieve such visualization, we manipulated individual, fluorescently labelled DNA molecules by using an optical trap, and detected their unwinding by RecBCD enzyme. DNA substrates were constructed by attaching a biotinylated oligonucleotide to one cohesive end of lambda (λ) DNA. These DNA molecules were attached at low density to 1- μm streptavidin-coated, polystyrene beads. The fluorescent dye YOYO-1 was bound to the DNA, and then RecBCD enzyme was bound to these fluorescent DNA molecules in the absence of ATP. Under these conditions, RecBCD enzyme binds only to the free end of the dsDNA, and neither translocates nor unwinds the DNA until ATP is introduced⁸.

These helicase–DNA complexes were introduced into one channel of a Y-shaped, micro-machined flow cell (Fig. 1a; and ref. 6). Helicase reaction buffer containing ATP was introduced into the second channel under conditions of laminar flow, creating a situation in

which the two solutions flowed parallel to one another with negligible mixing. The optically trapped bead appeared as a white sphere (the result of nonspecific binding of YOYO-1 to the bead), with an attached fluorescent DNA 'string' caused by flow-induced extension of the DNA molecule away from the trapped bead^{5,6}. Once a RecBCD enzyme–DNA–bead complex was trapped, the stage was moved to reposition this complex, from the sample side of the flow cell across the flow boundary between solutions to the reaction side. If a RecBCD enzyme molecule is attached at the end opposite to the bead, then DNA unwinding will proceed from that end towards the bead, resulting in a decreased length of duplex DNA owing to concomitant unwinding of the dsDNA and displacement of YOYO-1 molecules by a single translocating RecBCD enzyme; the displaced free dye molecules possess negligible solution fluorescence (Fig. 1).

The observed unwinding of individual DNA molecules is shown in Fig. 2. Sequential video frames from representative reactions in the presence and absence of ATP are shown in Fig. 2a and b, respectively (see Supplementary Information). Before movement of the bead with the attached RecBCD–DNA complex to the reaction (ATP) side of the flow cell, the length of the DNA remains constant (first ~18 s of Fig. 2c). On relocation to the reaction side of the flow cell, a rapid decrease ($0.145 \mu\text{m s}^{-1}$) in the DNA length is observed when 1 mM ATP is present. DNA shortening is linear and constant, with no discernible pauses for a period of ~56 s, at which point the change in DNA length stops abruptly. (The occasional instantaneous deviation of DNA length from the average linear behaviour arises from movement of DNA slightly in and out of the focal plane; similar behaviour is observed for experiments either with or without ATP). The length remains constant for the remaining 42 s of this time course, suggesting that the cessation of unwinding results from dissociation of RecBCD enzyme from the remaining dsDNA.

The observed rate of DNA shortening (from $t = 18$ s to $t = 74$ s) corresponds to an apparent rate of DNA helicase movement of

471 ± 30 base pairs (bp) s^{-1} (see Methods). The change in length of dsDNA that occurs before RecBCD enzyme dissociates, $8.3 \mu\text{m}$, corresponds to $26,800 \pm 700$ bp unwound by the single RecBCD enzyme. In contrast, in the absence of ATP there is only a small decrease in the apparent length of the DNA ($\sim 1 \mu\text{m}$) at a rate that is 16-fold lower than the ATP-dependent reaction ($-0.009 \mu\text{m s}^{-1}$ versus $-0.145 \mu\text{m s}^{-1}$). This change in apparent length is enzyme-independent, and results from a slow dissociation of YOYO-1 from sites across the entire length of the DNA molecule that occurs when the DNA molecule is relocated from the sample to the reaction side of the flow cell, owing to the lower YOYO-1 concentration on the reaction side (Fig. 2b; and data not shown). Control experiments done in the presence of ATP, but in the absence of enzyme, produced results identical to those obtained in the absence of ATP (data not shown) and showed that, at longer times, a new equilibrium DNA length was achieved (YOYO-1 elongates dsDNA by 1.1–1.2-fold; refs 5, 6, 9). When the unwinding rate in the presence of 1 mM ATP is corrected for the enzyme-free dissociation of YOYO-1, the corrected unwinding rate ($k_{\text{cat}}^{\text{corr}}$) is $443 \pm 30 \text{ bp s}^{-1}$. We conclude that this unwinding rate is due to a single molecule of RecBCD enzyme that bound to the free dsDNA end opposite the bead, and both translocated and unwound the DNA in an ATP-dependent manner once the DNA entered the ATP channel. Additional RecBCD enzyme molecules could not have contributed to the unwinding because RecBCD enzyme binds only to blunt, or nearly blunt dsDNA ends, with a greater than 1 million-fold preference relative to internal sites^{10,11}, and RecBCD enzyme was present only in the sample side of the flow cell.

To confirm further that the unwinding observed was dependent on RecBCD enzyme, we carried out assays at four additional ATP concentrations and also at 37°C , because previous steady-state kinetic assays showed that the rate of unwinding depends on both of these variables^{11,12}. As expected, the observed rate of unwinding increased with increasing ATP concentration (Fig. 3). In all cases,

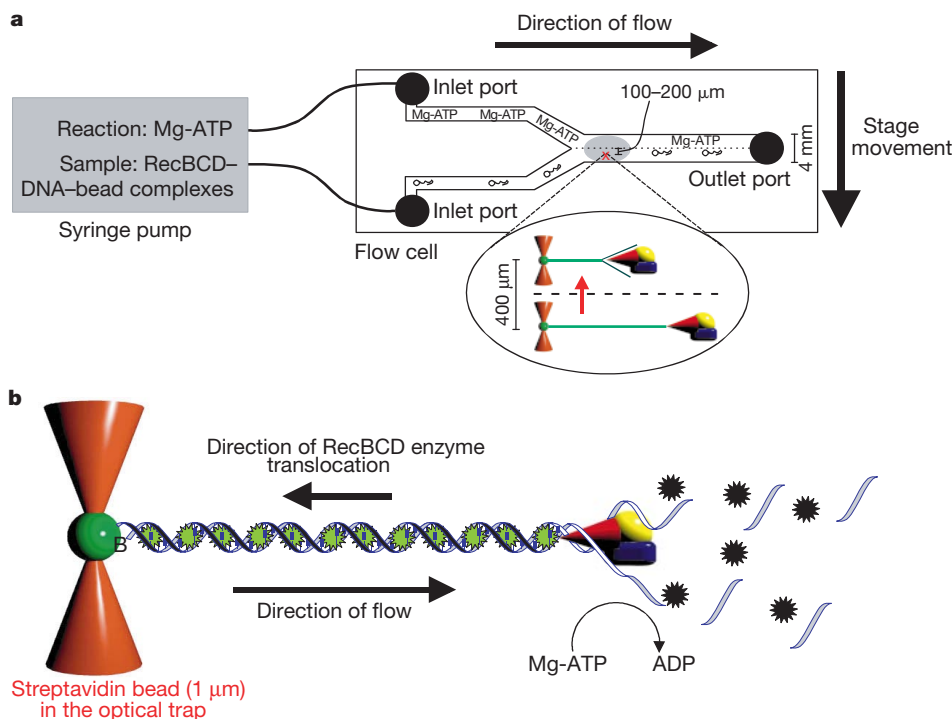


Figure 1 Visualization of DNA helicase action on individual DNA molecules. **a**, Syringe pump and flow cell: the sample syringe contains helicase–DNA–bead complexes, and the reaction syringe contains ATP. 'X' indicates the laser trap position, and the red arrow indicates movement of the trapped DNA–bead complex across the boundary between solutions. Inset, the trapped DNA with bound helicase, and its unwinding after relocation

into the reaction solution. **b**, Fluorescent DNA helicase assay⁷. A trapped and stretched, fluorescent DNA molecule is shown. As RecBCD enzyme translocates, it both unwinds and degrades the DNA, simultaneously displacing dye molecules (black stars). B, biotinylated oligonucleotide.

unwinding of dsDNA molecules was continuous and without any detectable pausing, indicating that the helicase action of individual RecBCD enzymes was not affected detectably by any sequence present in the λ DNA. Microscopic pausing, at the individual base-pair level, would not be detected in our experiments because this is beyond the temporal (33 ms per frame; hence pauses shorter than about 10 frames, or ~ 330 ms, would be difficult to detect) and spatial resolution of the assay system (a maximum of ~ 244 nm or 800 bp but, owing to occasional movement of the DNA out of the focal plane, the actual resolution is lower, ~ 1 μ m or $\sim 3,000$ bp).

Hence, there is no class of limited (10–20) specific pause sites at which the enzyme pauses for more than a fraction of a second or so.

Although the rate of DNA unwinding by any individual RecBCD enzyme molecule was uniform (within experimental error) on any given DNA molecule (Fig. 2), the rate for different helicase molecules deviated by 1.4–5-fold at each ATP concentration examined (Fig. 3). However, their average behaviour at each ATP concentration was similar to that observed in both steady-state solution experiments^{1,11} and electron microscopic assays that analysed intermediates of the unwinding reaction¹³. For example, at 250 μ M ATP

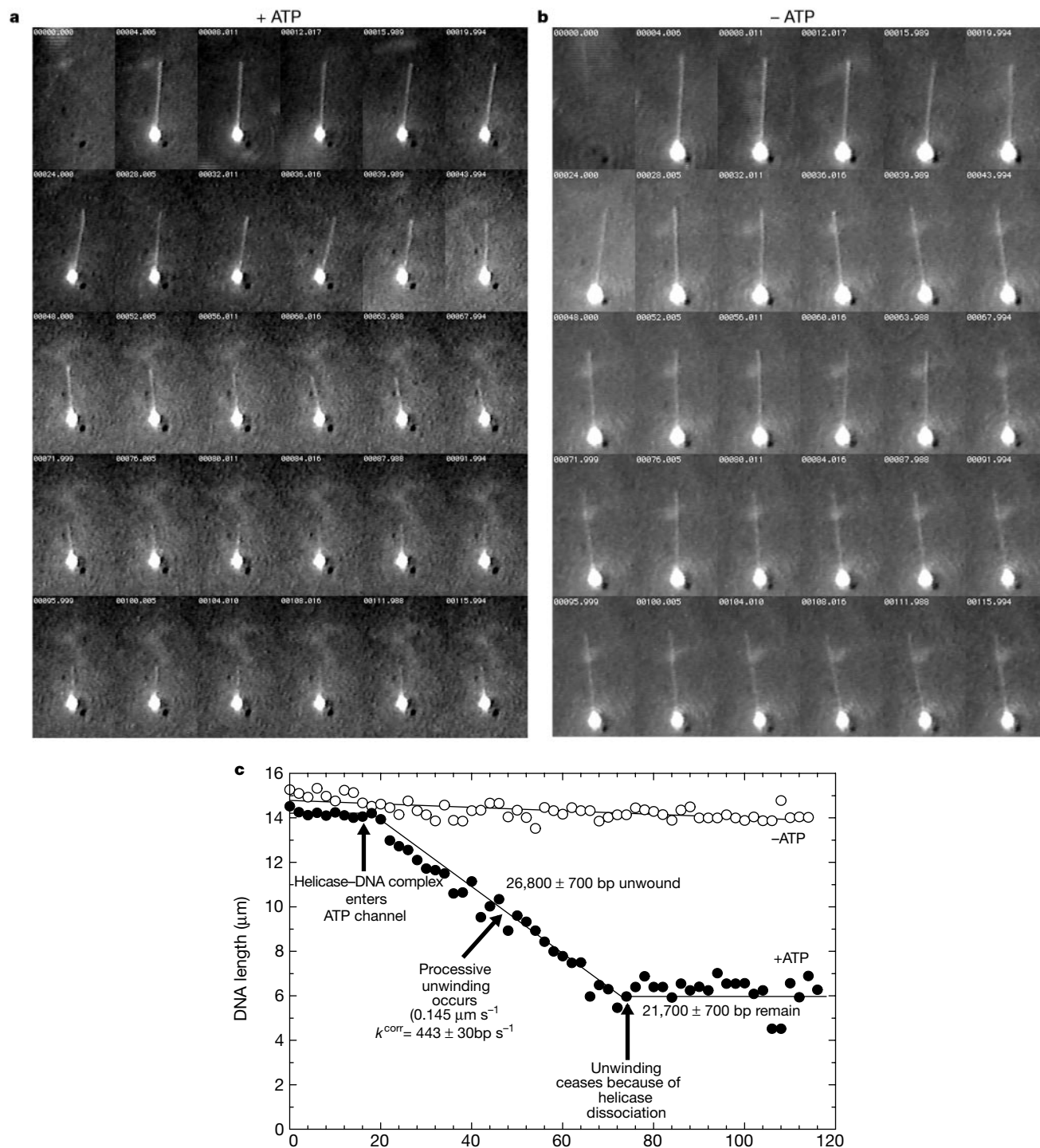


Figure 2 Unwinding of a DNA molecule by RecBCD enzyme. **a**, **b**, Selected, sequential frames from a video recording of reactions either in the presence (**a**) or absence (**b**) of ATP (1 mM). The direction of translocation and DNA unwinding by RecBCD enzyme is from the

DNA end opposite the bead, towards the bead (that is, from the top of each frame towards the bottom). Numbers at the top of each frame indicate elapsed time. **c**, Analysis of the time courses in **a** and **b**.

the unwinding rate for eight individual enzymes varied from a minimum of 156 bp s^{-1} to a maximum of 428 bp s^{-1} , with an average of $359 \pm 89 \text{ bp s}^{-1}$. A similar behaviour was observed at each of the four additional ATP concentrations used. Both this variation in unwinding rate by individual helicase molecules and the unfaltering movement could not have been discerned from steady-state experiments. At present, the source of this variation is unknown; some of the variation might be day-to-day variation or variation in the DNA–bead preparation used, but at least a twofold variance was observed for six molecules on the same day with the same DNA–bead preparation. Variation in individual enzymatic activity was observed for at least two other enzymes: lactate dehydrogenase¹⁴ and T7 DNA polymerase¹⁵. Despite the individual variation, when the averaged data from the unwinding of 42 DNA molecules (Fig. 3a) were fitted to the Michaelis–Menten equation, a V_{max} of $521 \pm 60 \text{ bp s}^{-1}$ and a $K_{\text{m}}^{\text{ATP}}$ of $142 \pm 58 \mu\text{M}$ was obtained. Both values are virtually identical to those reported from bulk-solution experiments ($586 \pm 45 \text{ bp s}^{-1}$ and $130 \pm 30 \mu\text{M}$ ATP, respectively^{11,12,16}). In addition to being dependent on ATP concentration, the rate of unwinding in 1 mM ATP increased twofold when the temperature was raised to 37°C (Table 1); this $k_{\text{cat}}^{\text{corr}}$ ($972 \pm 172 \text{ bp s}^{-1}$) is essentially identical to that obtained from steady-state measurements ($930 \pm 15 \text{ bp s}^{-1}$; ref. 11).

In addition to obtaining dsDNA unwinding rates, we directly observed the distance travelled by an individual RecBCD enzyme from its binding site at a dsDNA end to its point of dissociation, a distance that directly defines the processivity of DNA unwinding (that is, the number of base pairs unwound by a single helicase

Table 1 Summary of DNA helicase behaviour for individual RecBCD enzymes

Temperature ($^\circ\text{C}$)	ATP concentration (mM)	$k_{\text{cat}}^{\text{corr}}$ (bp s^{-1} per RecBCD)*	N (bp per end)†
23	0.05	161 ± 23 (5)‡	$6,700 \pm 4,300$
23	0.15	224 ± 166 (4)	$9,600 \pm 3,000$
23	0.25	359 ± 89 (8)	$22,900 \pm 9,500$
23	0.60	386 ± 71 (10)	$21,300 \pm 8,500$
23	1.00	502 ± 243 (5)	$27,000 \pm 9,800$
37	1.00	972 ± 172 (3)	$38,000 \pm 5,700$

*The observed unwinding rate as calculated for the linear portion of plots such as those shown in Fig. 2c. As the rate is determined from an individual enzyme molecule, we refer to it as k_{cat} ; it has been corrected for the ATP-independent rate, thus $k_{\text{cat}}^{\text{corr}}$. The observed variation is the standard deviation of unwinding rates for individual RecBCD enzyme molecules at that ATP concentration. † N is the number of base pairs of dsDNA translocated and unwound by an individual RecBCD enzyme molecule. The reported variation is the standard deviation of the processivity for all enzyme molecules at a particular ATP concentration.

‡The number in parentheses indicates the number of DNA molecules used to determine the unwinding rates and the processivities at each ATP concentration; eight molecules were examined in the absence of ATP.

molecule per DNA binding event before dissociation) (Fig. 3b and Table 1). The processivity varied in an ATP-dependent fashion, with the maximum value observed at 1 mM ATP. There was a stochastic variation in the processivity of individual RecBCD enzyme molecules at any given ATP concentration (Fig. 3b). For example, at $250 \mu\text{M}$ ATP, unwinding terminated abruptly after values ranging from 9,200 to 39,500 bp of dsDNA unwound by eight different RecBCD enzymes. The average value is $22,900 \pm 9,500 \text{ bp}$, which is similar to that determined previously ($27,000 \pm 3,000 \text{ bp}$; ref. 1).

We fitted the processivity data from the same 40 molecules analysed in Fig. 3a to a hyperbolic function to determine a maximum processivity at 23°C of $29,670 \pm 4,256 \text{ bp}$ per binding event, with an apparent $K_{\text{m}}^{\text{ATP}}$ for processive unwinding ($K_{\text{N}}^{\text{ATP}}$) of $158 \pm 78 \mu\text{M}$ (Fig. 3b; and ref. 1). These results are similar to those reported previously for RecBCD enzyme using both gel and fluorescence measurements ($32,000 \pm 1,800 \text{ bp}$ per end, and $K_{\text{N}}^{\text{ATP}} = 41 \pm 9 \mu\text{M}$; ref. 1). The processivity is even higher at elevated temperature, increasing to an average of $38,000 \pm 5,700 \text{ bp}$ per dsDNA end-binding event (Table 1); in fact one RecBCD enzyme molecule could unwind as much as 42,300 bp of dsDNA. As the translocation step size is 23 bp (ref. 17), a single RecBCD enzyme has the capacity to take as many as 1,840 steps (translocating a total distance of $13 \mu\text{m}$) before dissociating when the ATP concentration is not limiting.

We have observed both the rate and processivity of dsDNA unwinding for individual RecBCD enzyme molecules using a single-molecule helicase assay. To our knowledge, this is the first time that the action of individual DNA helicase molecules on dsDNA substrates has been observed in real time. The results obtained are consistent with and in precise agreement with previously published reports^{1,11,7,12}. This assay has further potential and can be used to study a variety of individual nucleic acid enzymes: both those that simply bind to nucleic acids to effect a conformational change (for example, the DNA strand-exchange protein, RecA protein, that displaces fluorescent dye molecules upon binding to dsDNA¹⁸), and those that process DNA in a manner similar to that observed here. □

Methods

Optical trapping and fluorescence microscopy

DNA helicase reactions were performed in a two-channel, Y-shaped, micro-machined glass flow cell⁶ (Fig. 1) with the inlet ports connected to 1-ml syringes controlled by a syringe pump (Model KDS 200; KD Scientific, Boston, MA). The flow cell was held in place on a motorized stage controlled by an MSI-2000 computerized stage controller ($0.1\text{-}\mu\text{m}$ resolution; Applied Scientific Instruments). The stage was positioned in a Nikon ES400 microscope, which was equipped for epifluorescence and modified to incorporate an optical trap. The optical trap used an Nd:YLF infrared laser (wavelength $1,047 \text{ nm}$, 500 mW ; Spectra Physics) focused through an oil-immersion objective lens (Plan Fluor $100\times$, 1.3 N.A. ; Nikon) and Immersol (518F, low fluorescence; a gift from Zeiss), to a position $10\text{--}15 \mu\text{m}$ below the upper surface of the flow cell. The flow cell is $\sim 4,000 \mu\text{m}$ wide, and trapping is initially done $100\text{--}200 \mu\text{m}$ from the boundary between solutions on the sample side of the flow cell, to ensure that a trapped complex that may have a RecBCD enzyme molecule attached, has not been exposed to ATP. The trapped complex is moved

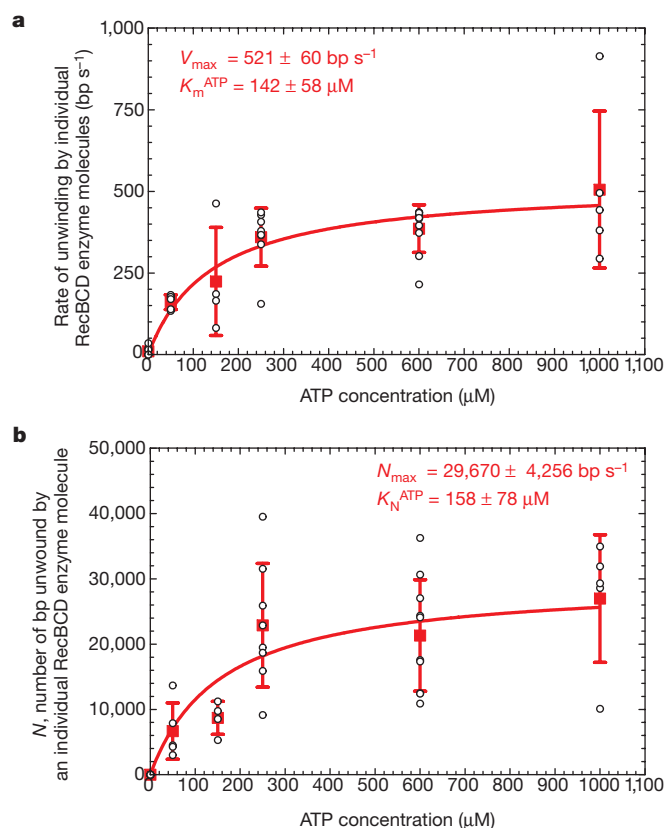


Figure 3 Both the rate and processivity for dsDNA unwinding by individual RecBCD enzyme molecules may vary, but their averages fall within ranges observed for bulk solution. For each ATP concentration (23°C), 4–10 DNA molecules were measured on different days, using several different preparations. **a**, Unwinding rates for single RecBCD enzyme molecules; **b**, processivities for the same molecules. Open circles represent individual molecule results, and squares represent their average for a given ATP concentration (error bars indicate the standard deviation). Data were fitted to a hyperbola (red line) for comparison with previously published steady-state values¹.

~400 μm into the reaction side to ensure that the reaction occurs in a homogenous concentration of ATP.

Fluorescent DNA–bead complexes were excited with the microscope's high-pressure mercury lamp using the appropriate filter set (blue filter set #11001; Chroma Technology, Brattleboro, VT). Fluorescence images were captured by a charge-coupled device camera (CCD-300T-IGF; Dage-MTI, Michigan City, IN) coupled to an image intensifier (VS4-1845; Video-Scope International, Sterling, VA), and were recorded on VHS videotape. To increase temperature to 37 °C, the microscope stage was enclosed in a Plexiglas housing and warm air was introduced.

DNA–bead preparation

The protocol used was modified from ref. 6. Bacteriophage λ DNA (0.096 pmol; New England Biolabs) was biotinylated at one end by annealing and ligating a 3'-biotinylated, 12-mer oligonucleotide (50 pmol; Operon Technologies) complementary to one of the cohesive ends. The biotinylated λ DNA ($3.6\text{--}7.2 \times 10^8$ molecules total) was reacted with 1 μm , streptavidin-coated, polystyrene beads (1.92×10^8 beads total; Bangs Laboratories) in 82 mM NaHCO_3 (pH 8.0) at 37 °C for 60 min. DNA–bead complexes were immediately transferred to a degassed solution containing 57 mM NaHCO_3 (pH 8.0), 30 mM dithiothreitol (DTT), 20% sucrose and 0.2 μM YOYO-1 (Molecular Probes). Dye binding was performed for a minimum of 60 min at 24 °C in the dark.

Single-molecule DNA helicase reactions

Before use, the flow cell was coated with either casein or BSA (100 $\mu\text{g ml}^{-1}$). The excess, unbound protein was washed out using 100 mM NaHCO_3 (pH 8.0). For each assay the sample syringe contained 41 mM NaHCO_3 buffer (pH 8.0), 35 mM DTT, 13% sucrose, 0.133 μM YOYO-1, 2 mM magnesium acetate, 1.92×10^8 DNA–bead complexes, and 4.6 nM RecBCD enzyme. The reaction syringe contained 41 mM NaHCO_3 buffer (pH 8.0), 35 mM DTT, 13% sucrose, 0.02 μM YOYO-1, 2 mM magnesium acetate, and ATP at the correct concentration. The *Escherichia coli* single-stranded DNA (ssDNA)-binding protein was not needed in these reactions because endonucleolytic cleavage of unwound ssDNA by the associated nuclease activity of RecBCD enzyme, releases the ssDNA as fragments that are immediately washed away in the buffer flow and, thus, do not accumulate to either inhibit the enzyme¹⁰ or affect the observed fluorescence signal⁷.

The RecBCD enzyme preparation used in all experiments was 100% active (data not shown) as determined using a spectrofluorometric helicase assay¹¹. Standard visualization reactions were done at room temperature (~ 23 °C) using degassed buffers. Flow was initially at 0.8 ml h^{-1} for 10 min and was gradually decreased in increments of 50% to a final flow rate 10–20 $\mu\text{l h}^{-1}$ (linear flow rates of $\sim 80\text{--}150$ $\mu\text{m s}^{-1}$), over a period of ~ 20 min.

Data analysis

Images were captured on a Power PC Macintosh computer interfaced with the VCR through an LG-3 frame-grabber card, operating at 1 frame per 33 ms, and controlled by Scion NIH Image v1.62c (Scion Corporation). Captured videos were converted into individual, time-stamped, sequential frames (Fig. 2) to make measurements. Individual DNA molecules were measured using the linear measurement tool of Scion NIH Image; calibration of the microscope optics was achieved using an Objective Micrometer (Fisher Scientific) that was marked in units of 10 μm .

The rate of DNA unwinding was calculated by measuring the observed length of DNA in each frame of a time course and fitting the resultant data to a linear function by least squares analysis. The resulting slope in each ATP-dependent reaction was corrected using the slope determined in the absence of ATP; this corrected slope was multiplied by the number of bp of λ DNA per μm to calculate the corrected rate in units of bp s^{-1} . Under the assay conditions used, the average length of individual, stretched, fluorescent λ DNA molecules was 14.9 μm ($n = 43$), which is shorter than previously published lengths^{5,6}. Those reports used a higher concentration of sucrose and were done in the absence of magnesium ions. In our experiments, the length of a single, stretched λ DNA molecule was affected by flow rate and by the concentrations of sucrose, magnesium ions, and YOYO-1, varying from more than 18 μm in 20% sucrose in the absence of magnesium acetate, to 14.9 μm in 13% sucrose and 2 mM magnesium acetate (data not shown). Thus, for all calculations, as λ DNA is 48,502 bp in length, and the observed average length of a λ DNA molecule is 14.9 μm , the average number of bp per μm is $48,502/14.9 = 3,255$.

Received 23 May; accepted 23 October 2000.

- Roman, L. J., Eggleston, A. K. & Kowalczykowski, S. C. Processivity of the DNA helicase activity of *Escherichia coli* recBCD enzyme. *J. Biol. Chem.* **267**, 4207–4214 (1992).
- Arnold, D. A. & Kowalczykowski, S. C. in *Encyclopedia of Life Sciences* [online] (Nature Publishing Group, London, 1999) (<http://www.els.net>).
- Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. & Rehauer, W. M. Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol. Rev.* **58**, 401–465 (1994).
- Kuzminov, A. Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. *Microbiol. Mol. Biol. Rev.* **63**, 751–813 (1999).
- Perkins, T. T., Smith, D. E. & Chu, S. Direct observation of tube-like motion of a single polymer chain. *Science* **264**, 819–822 (1994).
- Brewer, L. R., Corzett, M. & Balhorn, R. Protamine-induced condensation and decondensation of the same DNA molecule. *Science* **286**, 120–123 (1999).
- Eggleston, A. K., Rahim, N. A. & Kowalczykowski, S. C. A helicase assay based on the displacement of fluorescent, nucleic acid-binding ligands. *Nucleic Acids Res.* **24**, 1179–1186 (1996).
- Ganesan, S. & Smith, G. R. Strand-specific binding to duplex DNA ends by the subunits of *Escherichia coli* recBCD enzyme. *J. Mol. Biol.* **229**, 67–78 (1993).
- Bennink, M. L. *et al.* Single-molecule manipulation of double-stranded DNA using optical tweezers: interaction studies of DNA with RecA and YOYO-1. *Cytometry* **36**, 200–208 (1999).

- Taylor, A. F. & Smith, G. R. Substrate specificity of the DNA unwinding activity of the RecBC enzyme of *Escherichia coli*. *J. Mol. Biol.* **185**, 431–443 (1985).
- Roman, L. J. & Kowalczykowski, S. C. Characterization of the helicase activity of the *Escherichia coli* RecBCD enzyme using a novel helicase assay. *Biochemistry* **28**, 2863–2873 (1989).
- Eggleston, A. K. & Kowalczykowski, S. C. The mutant recBCD enzyme, recB¹⁰⁹CD enzyme, has helicase activity but does not promote efficient joint molecule formation *in vitro*. *J. Mol. Biol.* **231**, 621–633 (1993).
- Taylor, A. & Smith, G. R. Unwinding and rewinding of DNA by the recBC enzyme. *Cell* **22**, 447–457 (1980).
- Xue, Q. F. & Yeung, E. S. Differences in the chemical reactivity of individual molecules of an enzyme. *Nature* **373**, 681–683 (1995).
- Wuite, G. J. L., Smith, S. B., Young, M., Keller, D. & Bustamante, C. Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. *Nature* **404**, 103–106 (2000).
- Roman, L. J. & Kowalczykowski, S. C. Characterization of the adenosinetriphosphatase activity of the *Escherichia coli* RecBCD enzyme: Relationship of ATP hydrolysis to the unwinding of duplex DNA. *Biochemistry* **28**, 2873–2881 (1989).
- Bianco, P. R. & Kowalczykowski, S. C. Step size measurements on the translocation mechanism of the RecBC DNA helicase. *Nature* **405**, 368–372 (2000).
- Zaitsev, E. N. & Kowalczykowski, S. C. Binding of double-stranded DNA by *Escherichia coli* RecA protein monitored by a fluorescent dye displacement assay. *Nucleic Acids Res.* **26**, 650–654 (1998).
- Anderson, D. G. & Kowalczykowski, S. C. SSB protein controls RecBCD enzyme nuclease activity during unwinding: a new role for looped intermediates. *J. Mol. Biol.* **282**, 275–285 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We would like to thank S. Chan and J. Lengyel for assistance with measurements, and the following people for their comments on the manuscript: N. Handa, J. Kleiman, A. Mazin, J. New, E. Seitz, M. Spies, T. Sugiyama and Y. Wu. This work was supported by an NIH Grant to S.C.K. and a DOE Center of Excellence for Laser Applications in Medicine Grant to Y.Y. and R.J.B.

Correspondence and requests for materials should be addressed to S.C.K. (e-mail: skowalczykowski@ucdavis.edu).

Crystal structure of the transcription activator BmrR bound to DNA and a drug

Ekaterina E. Zhelezнова Heldwein & Richard G. Brennan

Department of Biochemistry and Molecular Biology, Oregon Health Sciences University, Portland, Oregon 97201-3098, USA

The efflux of chemically diverse drugs by multidrug transporters that span the membrane¹ is one mechanism of multidrug resistance in bacteria. The concentrations of many of these transporters are controlled by transcription regulators, such as BmrR in *Bacillus subtilis*², EmrR in *Escherichia coli*³ and QacR in *Staphylococcus aureus*⁴. These proteins promote transporter gene expression when they bind toxic compounds. BmrR activates transcription of the multidrug transporter gene, *bmr*, in response to cellular invasion by certain lipophilic cationic compounds (drugs)^{2,5,6}. BmrR belongs to the MerR family, which regulates response to stress such as exposure to toxic compounds or oxygen radicals in bacteria^{7–12}. MerR proteins have homologous amino-terminal DNA-binding domains but different carboxy-terminal domains, which enable them to bind specific 'coactivator' molecules. When bound to coactivator, MerR proteins upregulate transcription by reconfiguring the 19-base-pair spacer found between the –35 and –10 promoter elements to allow productive interaction with RNA polymerase^{7,9–12}. Here we report the 3.0 Å resolution structure of BmrR in complex with the drug tetraphenylphosphonium (TPP) and a 22-base-pair oligodeoxynucleotide encompassing the *bmr* promoter. The structure reveals an unexpected mechanism for transcription activation that involves localized base-pair breaking, and base sliding and realignment of the –35 and –10 operator elements.

The structure of the BmrR–TPP–*bmr* promoter complex was determined by a combination of molecular replacement and multiple isomorphous replacement (Supplementary Information, Table 1). The asymmetric unit contains one BmrR monomer and a promoter half-site, which results in the statistical disorder of several base pairs of this pseudo-palindrome (Supplementary Information, Fig. 1). The monomer (Fig. 1a) contains three domains: the N-terminal DNA-binding domain (residues 1–75); the linker, containing an 11-turn α -helix, $\alpha 5$ (residues 76–119), that connects the N- and C-terminal domains; and the C-terminal drug-binding domain, the so-called BRC, (residues 120–278). The DNA-binding domain belongs to the superfamily of winged-helix proteins and contains a four-helix bundle and a three-stranded antiparallel β -sheet¹³. The topology is $\beta 1$ – $\alpha 1$ – $\alpha 2$ – $\beta 2$ – $\beta 3$ – $\alpha 3$ – $\alpha 4$. Conserved hydrophobic core residues throughout the MerR family indicate that their DNA-binding domains assume similar folds (Supplementary Information, Fig. 2). The structure of the drug-binding domain, a distorted eight-stranded β -barrel, is identical to the previously solved structure of the individually expressed apo and TPP-bound BRC¹⁴ (Supplementary Information).

The structure of dimeric BmrR bound to DNA resembles a butterfly (Fig. 1b). Each domain participates in dimerization, burying $\sim 5,800 \text{ \AA}^2$ of accessible surface area¹⁵. The DNA-binding domain of each monomer packs tightly against the drug-binding domain of its dimerization partner, and buries $\sim 2,000 \text{ \AA}^2$ of accessible surface area (ASA) per monomer (Fig. 1b, c). Helix $\alpha 6$ packs against helix $\alpha 3'$ and the C terminus of helix $\alpha 1'$, where the prime indicates the other subunit. In addition, the loop connecting strands $\beta 10$ and $\beta 11$ wedges between helices $\alpha 3'$ and $\alpha 4'$ of the DNA-binding domain. Linker helices $\alpha 5$ and $\alpha 5'$ bury $\sim 900 \text{ \AA}^2$ of ASA per monomer by forming an antiparallel coiled coil (Fig. 1b), which is likely to be found in all MerR family proteins¹⁶. Helix $\alpha 5$ also makes van der Waals contacts to helix $\alpha 3'$ and the loop between strands $\beta 10'$ and $\beta 11'$.

BmrR uses a helix–turn–helix (HTH) motif and two ‘wings’, W1 and W2, to bind the *bmr* promoter (Figs 1b and 2a). The HTH motif consists of helices $\alpha 1$ and $\alpha 2$, and the turn connecting them. Overlays of the 21-residue HTH domains of BmrR and of bacterial HTH proteins, PurR and TetR, as well as of bacterial winged-helix proteins, CAP and BirA, result in r.m.s. deviations of 0.9 Å, 1.1 Å, 1.2 Å and 1.2 Å, respectively. The symmetry-related recognition helices (residues 19–28) contact two consecutive major grooves whereby each helical axis is nearly perpendicular to the local DNA helical axis. Although helix $\alpha 1$ contributes a hydrogen bond from

the amide of Gly 9 to the phosphate of cytosine 9' and van der Waals contacts from the C $\gamma 2$ of Ile 8 to the deoxyribose ring of cytosine 9', most HTH contacts are made by residues from helix $\alpha 2$ (Fig. 2). Hydrogen bonds occur between the amide of Ala 21 and guanine 4 phosphate, the NH₂ and N ϵ of Arg23 and cytosine 8' phosphate, and the Tyr 25 hydroxyl group and guanine 3 phosphate. van der Waals contacts include those between the NH₂ of Arg 23 and the deoxyribose ring of cytosine 9' and C7 atom of thymine 7'; the NH1 of Arg 23 and C7 atom of thymine 7'; and the Tyr 24 phenyl ring and C8 atom of adenine 2 and deoxyribose rings of guanine 3 and adenine 2 (Fig. 2b). Ser 18 and Lys 20 probably make DNA contacts, but their side chains are disordered.

The second DNA-binding element, wing W1, comprises strands $\beta 2$ and $\beta 3$ and their connecting loop (residues 35–46). Specifically, the hydroxyl group of Ser 41 makes hydrogen bonds to the cytosine 8' phosphate, whereas that of Tyr 42 hydrogen bonds to the O4' atom of the cytosine 9' deoxyribose ring and the O2 of cytosine 10', and makes van der Waals contacts to the guanine 10 N2. NH1, NH2 and the backbone amide of Arg 43 make hydrogen bonds to the phosphates of thymine 7' and cytosine 8' and a van der Waals contact to the cytosine 8' deoxyribose ring. The N ϵ of Arg 43 makes a hydrogen bond to O $\delta 1$ of Asp 26, which positions the Arg 43 side chain for reading the DNA (Fig. 2a). Arg 43 and a glutamate, aspartate or glutamine at position 26 are conserved in the MerR family indicating their importance in DNA binding by W1.

Unlike other winged-helix proteins¹³, the third DNA-binding element of BmrR, wing W2, is not a loop, but rather another HTH motif composed of helices $\alpha 3$ and $\alpha 4$, and their connecting turn (Figs 1b and 2a). The $\alpha 3$ and $\alpha 4$ helices are less crossed than those of the major-groove binding HTH motif, and superimposition of their C α atoms (residues 9–26 and 56–73) results in an r.m.s. deviation of 1.9 Å. The protein–DNA contacts made by W2 include hydrogen bonds between the N ϵ of Lys 60 and guanine 3 phosphate and the NH of Leu 66 and the adenine 2 phosphate. The dipole moment of helix $\alpha 4$ also contributes to binding as its electropositive N terminus points directly at the adenine 2 phosphate. Despite the large number of contacts between BmrR and its promoter, the DNA remains readily accessible for binding by RNA polymerase⁷ (Fig. 1b).

The structure of the complex reveals no base-specific contacts. This might be due to the statistical disordering of the DNA or to potential specific contacts that are unresolved at the current resolution. BmrR might also achieve DNA recognition through

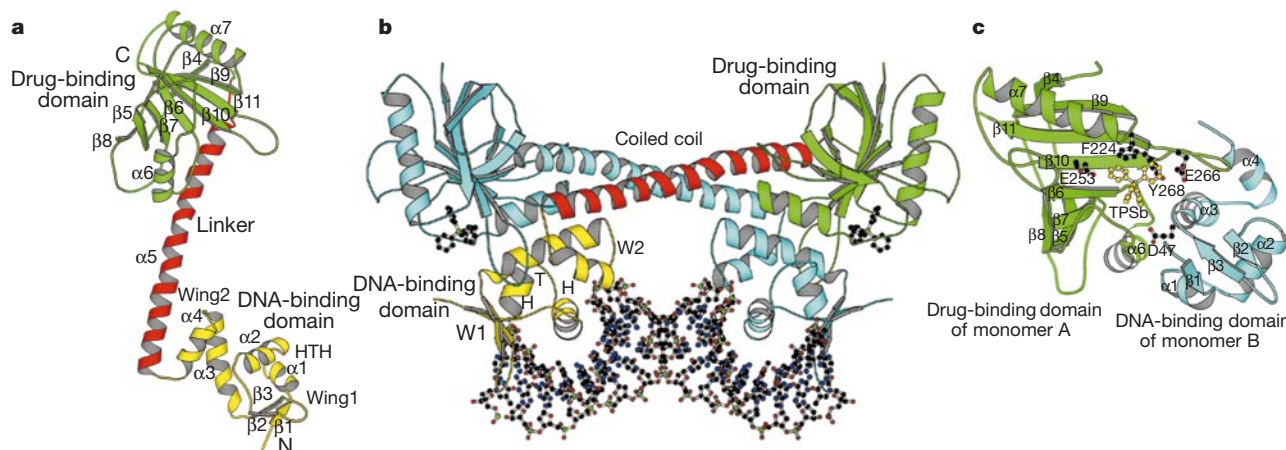


Figure 1 Crystal structure of the BmrR–drug–DNA complex. **a**, BmrR monomer. The DNA-binding domain, α -helical linker and drug-binding domain are shown in yellow, red and green, respectively. **b**, BmrR dimer bound to DNA. One monomer is coloured as in **a**, whereas the other monomer is shown in cyan. DNA and TPP/TPSb are represented as

balls and sticks (carbon, black; nitrogen, blue; oxygen, red; and phosphorus/antimony, green). **c**, Dimerization interface between the drug-binding domain of a BmrR monomer (green) and the DNA-binding domain of its dimeric mate (cyan). The TPP/TPSb molecule and selected drug-binding residues are represented as ball and sticks.

indirect readout. Alternatively, because BmrR is always bound to its promoter², perhaps only apo BmrR forms sequence-specific contacts with the *bmr* promoter. Drug binding would then enforce a different, high affinity protein–DNA interface characterized by the apparent lack of base-specific contacts.

The most striking feature of the BmrR–TPP–DNA complex is the DNA structure (Fig. 1b). Globally, the promoter is bent by $\sim 50^\circ$ at its middle towards the major groove and away from the protein; but many DNA-binding proteins bend and kink their operators. What makes the *bmr* promoter unique is the base-pair breaking and base sliding of the central base-pair step that is effected by the BmrR–TPP complex. In essence, the A–T base pairs that surround the pseudo dyad of the *bmr* promoter break, and the unpaired adenine and thymine slide away from each other towards the 3' direction (Fig. 3a; and Supplementary Information, Fig. 1). Their final positions redefine the pseudo dyad of the promoter. Neither base participates in base pairing (Fig. 3a), and, because of the super-

imposition of the DNA and crystallographic dyads, adenine 1 and thymine 1 are related by a two-fold axis. This results in their statistical disorder and, owing to limited resolution of the data and the better fit of a pyrimidine to the electron density, both were refined as thymine 1. The bases at positions 2 and 2', refined as adenine 2 and thymine 2', also undergo displacement but still form a distorted Watson–Crick base pair. The presence of unpaired bases, immediately adjacent to the dyad axis, creates ripples on both strands whereby thymine 1 and adenine 2 on either strand are positioned against thymine 2' on the opposite strand (Figs 2a and 3a; and Supplementary Information, Fig. 1).

The adjacent bases thymine 2' and thymine 1 are unstacked, rolled out and undertwisted with rise, roll and helical twist values of 6.0 Å, 30° and 23° , respectively¹⁷. As a result, the operator 'bunches up' in the middle. The distorted DNA conformation is stabilized by interactions between residues Tyr 24, Tyr 25, Lys 60 and the N terminus of helix $\alpha 4$ with the phosphate backbone (Fig. 2a, b).

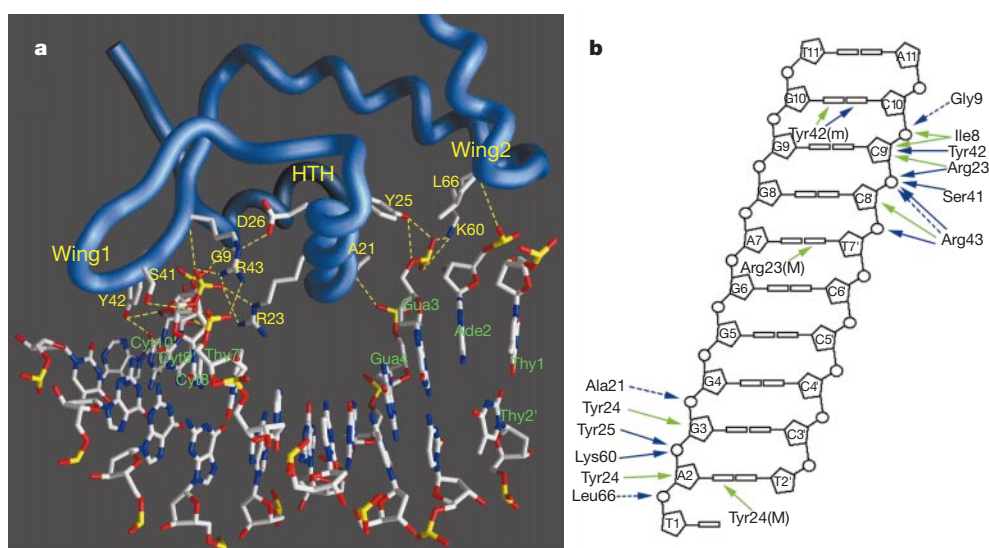


Figure 2 BmrR–DNA interactions. **a**, BmrR–DNA half-site interface. Atoms are represented as sticks (carbon, black; nitrogen, blue; oxygen, red; phosphorus, yellow). Hydrogen bonds are represented as yellow dashed lines. **b**, Representation of BmrR–DNA contacts. DNA is shown as a cylindrical projection where bases are depicted as rectangular boxes, deoxyribose rings as pentagons, and phosphates as circles. Hydrogen

bonds are represented as blue, and van der Waals interactions as green arrows. Solid and dashed lines represent contacts from side chains and backbone amides, respectively. Side chains that contact bases in the major or minor grooves are labelled with 'M' and 'm', respectively.

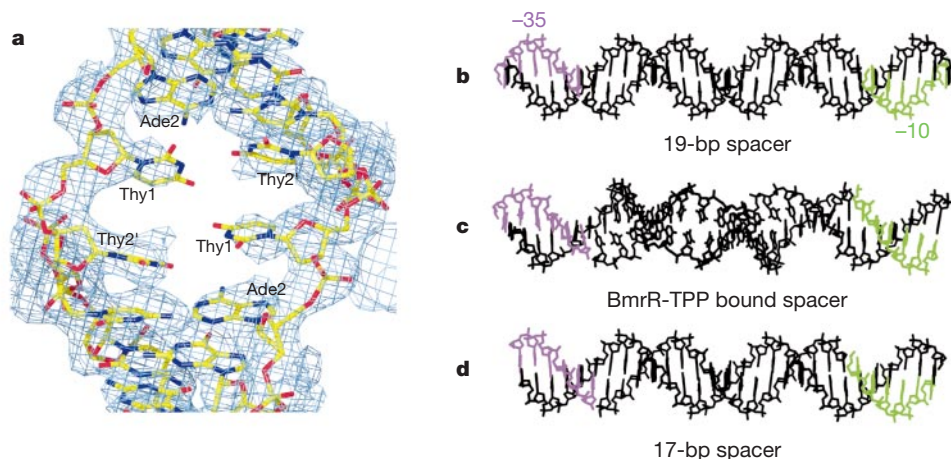


Figure 3 *bmr* promoter. **a**, The simulated-annealing omit electron density map at 1.0σ contour level, where Thy1, Thy2' and Ade2, and all atoms within a 5 Å radius were deleted. DNA atoms are represented as sticks (carbon, yellow; nitrogen, blue; oxygen, red; phosphorus, yellow). **b**, Promoter with a 19-bp spacer. **c**, Promoter with a 19-bp spacer

as observed in the BmrR–TPP bound operator. The full-length *bmr* promoter was modelled by adding 3-bp upstream and 7-bp downstream. **d**, Promoter with a 17-bp spacer. The –35 (TTGACT) and –10 (TACAGT) boxes of one strand are shown in pink and green, respectively.

A similar DNA structure is anticipated for all 'activated' MerR family members because each has a hydrogen-bond donor (lysine, arginine, glutamine) at position 60 and a completely conserved Tyr 25. Beyond this localized distortion, the remaining base pairs of the operator, 3/3' to 11/11', return to a duplex B-DNA like conformation with an average rise, roll and helical twist of 3.2 Å, 0.4°, and 32.0°, respectively. However, base pairs 6/6' to 10/10' are bent slightly (13°) towards the recognition helices (Figs 1b and 2a).

The mechanistic importance of the DNA distortion and altered conformation is underscored by previous biochemical studies on promoters regulated by MerR and SoxR, whereby deletion of 2 base pairs (bp) from their 19-bp spacers resulted in regulator-independent promoters^{18,19}. In a canonical B-DNA conformation, a 19-bp spacer places the -35 and -10 elements of the *bmr* promoter onto opposite sides of the DNA double helix, thereby preventing RNA polymerase from productively accessing both elements simultaneously (Fig. 3b). In contrast, a 17-bp spacer in the B-DNA conformation brings the -35 and -10 promoter elements onto the same side of the DNA (Fig. 3d). This spacer-length is found in most σ^A -regulated promoters. The result of the BmrR-TPP-induced base unpairing and sliding is to shorten the DNA by ~5 Å, or about 2 bp, compared to the end-to-end length of a 21-bp B-DNA. More importantly, the local untwisting of the DNA brings the -35 and -10 promoter elements onto the same face of the DNA double helix; much like the transcription-ready 17-bp spacer does (Fig. 3c, d).

To restructure the *bmr* promoter, DNA-bound BmrR must bind a coactivator—one of several structurally unrelated lipophilic cations, such as TPP^{2,5}. The present structure includes a molecule of tetraphenylantimonium (TPSb), a tetraphenylphosphonium analogue (Supplementary Information, Table 1). The drug-binding pocket is created primarily by residues from the drug-binding domain, with an unanticipated contribution from the DNA-binding domain (Fig. 1c). The phenyl rings of the TPSb molecule make van der Waals and stacking interactions with the side chains of residues Phe 224 and Tyr 268. Complementary electrostatic interactions are found between the TPSb molecule and carboxylate groups of Asp 47, Glu 253 and Glu 266 (Fig. 1c), and backbone carbonyl oxygens of loop residues Pro 144, Glu 145 and Asn 146. This binding mode, based on electrostatic, stereochemical and steric complementarity between the drug and the drug-binding pocket, is similar to that observed in the BRC-TPP structure¹⁴. However, the roles of Asp 47 and Glu 266 are unclear because BRC, which lacks Asp 47, and BmrR bind the same drugs with equal affinities⁵. Moreover, in the structure of the BRC-TPP complex, Glu 266 is disordered with only Glu 253 contributing to drug-protein charge complementarity.

In the absence of the structure of the apo BmrR-DNA complex, how drug binding is coupled to DNA distortion and transcription activation is unclear. Superimposition of residues 121–277 of the apo BRC and TPP-bound BmrR structures suggests that helix $\alpha 6$ of the drug-binding domain repacks against helix $\alpha 3'$ of the DNA-binding domain as a key part of the signalling mechanism. This repacking will clearly also impinge on helices $\alpha 4'$ and $\alpha 5$ (Fig. 1b). The signal of drug binding could be then transmitted in two ways: directly, toward the protein-DNA interface; and allosterically, towards the protein-DNA interface of the other subunit through the coiled coil, the latter of which is supported by mutational studies on MerR^{16,20}.

The structure of the BmrR-TPP-DNA complex provides an excellent model for understanding transcriptional regulation by other MerR family proteins. For example, the Hg²⁺-binding residues of MerR²¹ can be placed at the N terminus of helix $\alpha 5$ (Cys 82), near the C terminus of helix $\alpha 5'$ (Cys 117), and on a nearby loop (Cys 126), thus bringing their sulphhydryls into proximity for facile trigonal coordination of Hg²⁺. Activation-deficient mutants of SoxR, which binds 2Fe-2S redox centres, map to four cysteines of the C-terminal portion of $\alpha 5$ and slightly beyond²²

(Supplementary Information, Fig. 2). The constitutively active MerR mutants S86C and A89V can be rationalized^{16,20} as the result of a readily formed Cys 86/Cys 117 disulphide bond or new van der Waals interactions between A89V and its hydrophobic neighbours. Both would lock the apo protein into the activated form. Other constitutively active MerR phenotypes¹⁶ map to helices $\alpha 3$ and $\alpha 5$, and again can be explained readily by hydrogen bonds or van der Waals interactions that favour the activated conformation of the protein. By contrast, the activation-deficient MerR mutant⁷, A60V, would sterically prohibit the packing of helices $\alpha 3'$ and $\alpha 4'$ of the Hg²⁺-bound protein. The locations of these mutations underscore the critical role of the $\alpha 3'/\alpha 4'/\alpha 5$ interfaces in transcription regulation by the MerR family. □

Methods

Sample preparation and crystallization

The C-terminal His-tagged *bmrR* gene, subcloned into the pBAD plasmid under control of an arabinose-inducible promoter, was kindly provided by P. N. Markham and A. A. Neyfakh (University of Illinois, Chicago, Illinois). *E. coli* Top10 cells were transformed with this vector, grown at 37 °C to an absorbance at 595 nm of 0.6–0.8, and induced with 0.03% D-arabinose. After a 3-h induction, the cells were pelleted and lysed using a French press. BmrR was purified by Ni-NTA column chromatography and eluted with 400 mM imidazole. BmrR is poorly soluble and precipitates immediately after elution; however, the protein was resolubilized by adding a 22-bp DNA duplex encompassing the *bmr* operator sequence with a 5' A or T overhang (Supplementary Information, Fig. 1a). The BmrR-DNA mixture was dialysed against 20 mM Tris-HCl, pH 7.6, 0.2 M NaCl and 10% glycerol, and concentrated to 5–8 mg ml⁻¹. TPP, dissolved in dimethyl sulphoxide (DMSO), was added to the BmrR-DNA mixture to final concentrations of 5 mM TPP and 1% DMSO. Crystals of the ternary complex were grown at room temperature by the hanging-drop vapour-diffusion method against reservoirs containing 1.0–1.2 M imidazole, pH 6.5.

Data collection and structure determination

X-ray intensity data for the BmrR-TPP-DNA complex were collected at room temperature with an R-Axis IV imaging plate system, mounted on a Rigaku RU300 X-ray generator equipped with Yale focusing mirrors, and processed using BIOTEX (Molecular Structure Corporation, Woodlands, Texas). The crystals took the tetragonal space group *P*₄₁/2₂ with unit-cell dimensions: *a* = *b* = 109.5 Å, *c* = 144.3 Å, α = β = γ = 90°. The structure of the BmrR-TPP-DNA complex was solved by a combination of molecular replacement²³ and multiple isomorphous replacement²⁴. Heavy atom derivatives were prepared by soaking native crystals in solutions of K₂PtCl₄ or thimerosal, and by crystallizing the complex with 5-iodouridine-containing DNA (Supplementary Information, Fig. 1c). Initially, the location of the BRC domain was determined by molecular replacement using data in the resolution range 15.0–4.0 Å and the entire apo BRC monomer as a search model. Only space group *P*₄₃22 yielded a correct solution, which had a correlation coefficient of 45.2% and an *R*_{factor} of 51.9%.

Because no connected density was observed for the N-terminal domain or the DNA, phases were calculated from the molecular replacement solution to locate the single iodine site in the Iodo1 derivative. This allowed the identification of one platinum and one mercury site in subsequent cross-difference Fourier analyses. The heavy atom parameters (Supplementary Information) were refined by the maximum-likelihood method to 3.0 Å resolution²⁴. A solvent-flattened map (78% estimated solvent) revealed the density for the DNA-binding domain and the α -helical linker. A polyaniline model was built to the density using O²⁵. Phase combination²⁴ showed clear density for most of the BmrR residues and the DNA phosphate backbone. Intensity data sets were also collected on crystals of BmrR complexes in which the DNA contained 5-iodouridine at positions 2', 7', 1 and 9', referred to as complexes Iodo2, Iodo3, Iodo5 and Iodo6, respectively (Supplementary Information, Fig. 1). These data allowed the correct base assignment of the DNA sequence.

Before refinement, 5% of all data were set aside for cross-validation²⁶. Through iterative cycles of refinement with TNT²⁷ and model rebuilding²⁵, residues 3–277 and the DNA were built. A metal ion coordinated by His 189 was refined as a Zn²⁺ ion at occupancy of 0.5. The final model was verified by a series of *F*_o–*F*_c omit and simulated annealing omit electron density maps. Residues 1–2, 278, the six histidines of the C-terminal His-tag, and the DNA 5'-overhangs are disordered and not included in the final model. Forty-five amino-acid residues were refined as alanines because their side-chain density is either poor or missing. The final model has no Ramachandran outliers, as assessed by the program PROCHECK²⁸ (Supplementary Information, Table 1). The secondary structure elements were assigned by PROCHECK²⁸ and include β -strands $\beta 1$ (residues 5–7), $\beta 2$ (residues 34–36), $\beta 3$ (43–46), $\beta 4$ (124–128), $\beta 5$ (133–138), $\beta 6$ (170–174), $\beta 7$ (189–193), $\beta 8$ (208–211), $\beta 9$ (216–223), $\beta 10$ (248–257) and $\beta 11$ (268–276), and α -helices $\alpha 1$ (8–14), $\alpha 2$ (19–28), $\alpha 3$ (48–62), $\alpha 4$ (66–74), $\alpha 5$ (77–116), $\alpha 6$ (153–163) and $\alpha 7$ (226–242). Because of the poor density of the TPP molecule, isomorphous crystals of BmrR-TPSb-DNA complex were obtained by soaking TPSb into crystals of BmrR-TPP-DNA complex. The resulting difference electron-density map revealed the density for the entire drug. The BmrR-TPSb-DNA complex was refined with TNT²⁷.

Figures were prepared using MOLSCRIPT²⁹, GRASP¹⁵ and O²⁵.

Received 15 September; accepted 6 November 2000.

1. Saier, M. H. Jr *et al.* Evolutionary origins of multidrug and drug-specific efflux pumps in bacteria. *FASEB J.* **12**, 265–274 (1998).
2. Ahmed, M., Borsch, C. M., Taylor, S. S., Vazquez-Laslop, N. & Neyfakh, A. A. A protein that activates expression of a multidrug efflux transporter upon binding the transporter substrates. *J. Biol. Chem.* **269**, 28506–28513 (1994).
3. Brooun, A., Tomashek, J. J. & Lewis, K. Purification and ligand binding of EmrR, a regulator of a multidrug transporter. *J. Bacteriol.* **181**, 5131–5133 (1999).
4. Grkovic, S., Brown, M. H., Roberts, N. J., Paulsen, I. T. & Skurray, R. A. QacR is a repressor protein that regulates expression of the *Staphylococcus aureus* multidrug efflux pump QacA. *J. Biol. Chem.* **273**, 18665–18673 (1998).
5. Markham, P. N., LoGuidice, J. & Neyfakh, A. A. Broad ligand specificity of the transcriptional regulator of the *Bacillus subtilis* multidrug transporter Bmr. *Biochem. Biophys. Res. Commun.* **239**, 269–272 (1997).
6. Vazquez-Laslop, N., Markham, P. N. & Neyfakh, A. A. Mechanism of ligand recognition by BmrR, the multidrug-responsing transcriptional regulator: mutational analysis of the ligand-binding site. *Biochemistry* **38**, 16925–16931 (1999).
7. Summers, A. O. Untwist and shout: a heavy metal-responsive transcriptional regulator. *J. Bacteriol.* **174**, 3097–3101 (1992).
8. Holmes, D. J., Caso, J. L. & Thompson, C. J. Autogenous transcriptional activation of a thiostrepton-induced gene in *Streptomyces lividans*. *EMBO J.* **12**, 3183–3191 (1993).
9. Ansari, A. Z., Bradner, J. E. & O'Halloran, T. V. DNA-bend modulation in a repressor-to-activator switching mechanism. *Nature* **374**, 371–375 (1995).
10. Gaudy, P. & Weiss, B. SoxR, a [2Fe-2S] transcription factor, is active only in its oxidized form. *Proc. Natl Acad. Sci. USA* **93**, 10094–10098 (1996).
11. Hidalgo, E., Ding, H. & Dimple, B. Redox signal transduction via iron-sulfur clusters in the SoxR transcription activator. *Trends Biochem. Sci.* **22**, 207–210 (1997).
12. Outten, C. E., Outten, F. W. & O'Halloran, T. V. DNA distortion mechanism for transcriptional activation by ZntR, a Zn(II)-responsive MerR homologue in *Escherichia coli*. *J. Biol. Chem.* **274**, 37517–37524 (1999).
13. Gajiwala, K. S. & Burley, S. K. Winged helix proteins. *Curr. Opin. Struct. Biol.* **10**, 110–116 (2000).
14. Zhelezanova, E. E., Markham, P. N., Neyfakh, A. A. & Brennan, R. G. Structural basis of multidrug recognition by BmrR, a transcription activator of a multidrug transporter. *Cell* **96**, 353–362 (1999).
15. Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149 (1995).
16. Caguiat, J. J., Watson, A. L. & Summers, A. O. Cd(II)-responsive and constitutive mutants implicate a novel domain in MerR. *J. Bacteriol.* **181**, 3462–3471 (1999).
17. Lavery, R. & Sklenar, H. Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.* **6**, 655–667 (1989).
18. Parkhill, J. & Brown, N. L. Site-specific insertion and deletion mutants in the *mer* promoter-operator region of Tn501; the nineteen base-pair spacer is essential for normal induction of the promoter by MerR. *Nucleic Acids Res.* **18**, 5157–5162 (1990).
19. Hidalgo, E. & Dimple, B. Spacing of promoter elements regulates the basal expression of the *soxS* gene and converts SoxR from a transcriptional activator into a repressor. *EMBO J.* **16**, 1056–1065 (1997).
20. Comess, K. M., Shewchuk, L. M., Ivanetich, K. & Walsh, C. T. Construction of a synthetic gene for the metalloregulatory protein MerR and analysis of regionally mutated proteins for transcriptional regulation. *Biochemistry* **33**, 4175–4186 (1994).
21. Zeng, Q., Stalhandske, C., Anderson, M. C., Scott, R. A. & Summers, A. O. The core metal-recognition domain of MerR. *Biochemistry* **37**, 15885–15895 (1998).
22. Bradley, T. M., Hidalgo, E., Leautaud, V., Ding, H. & Dimple, B. Cysteine-to-alanine replacements in the *Escherichia coli* SoxR protein and the role of the [2Fe-2S] centers in transcriptional activation. *Nucleic Acids Res.* **25**, 1469–1475 (1997).
23. Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr. D* **55**, 484–491 (1999).
24. Furey, W. & Swaminathan, S. PHASES-95: a program package for the processing and analysis of diffraction data from macromolecules. *Methods Enzymol.* **277B**, 590–620 (1997).
25. Jones, T. Z., Zou, J. -Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
26. Brünger, A. T. Free R value: cross-validation in crystallography. *Methods Enzymol.* **277B**, 366–396 (1997).
27. Tronrud, D. E. TNT refinement package. *Methods Enzymol.* **277B**, 306–319 (1997).
28. Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
29. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

The authors thank P. N. Markham and A. A. Neyfakh for providing the *bmrR* expression construct and for helpful comments. This research was supported by the NSF, NIH and the

Medical Research Foundation of Oregon (R.G.B.) and the American Heart Association (E.E.Z.H.).

Correspondence and requests for materials should be addressed to R.G.B. (e-mail: brennanr@ohsu.edu). Coordinates for the BmrR–TPP–DNA and BmrR–TPSb–DNA complexes have been deposited in the RCSB Protein database under accession code *lexj* and *lexi*, respectively.

correction

Halocarbons produced by natural oxidation processes during degradation of organic matter

F. Keppler, R. Eiden, V. Niedan, J. Pracht & H. F. Schöler

Nature **403**, 298–301 (2000).

Throughout Figs 2–4 and the corresponding legends, mM should have been mmol, μ M should have been μ mol, and pM should have been pmol. In addition, in Fig. 3a the *x* axis should have been labelled 'Fe (II) conc. in medium (μ mol)'; and in Fig. 3b the *x* axis should have been labelled 'Halide ion conc. in medium (mmol)'. □

erratum

Memory B-cell persistence is independent of persisting immunizing antigen

Mitsuo Maruyama, Kong-Peng Lam & Klaus Rajewsky

Nature **407**, 636–642 (2000).

In this paper, Table 1 was printed incorrectly. It should have appeared as below. □

Table 1 *in vitro* secondary antibody responses of PE⁺IgG1⁺ memory B cells

Memory B cells	CD4 ⁺ T cells	IgG1	Antibodies (μ g ml ⁻¹)				
			Anti-PE		Anti-NP		
			IgM	IgA	IgG1	IgM	IgA
PE ⁺ NP ⁺	PE primed	34.1±1.0	<0.1	33.7±0.1	<0.1	<0.1	<0.1
PE ⁺ NP ⁺	NP/CG primed	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
NP ⁺ PE ⁺	PE primed	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1

Numbers represent the concentration of anti-PE or anti-NP antibody (in μ g ml⁻¹). Titres were obtained from three independent wells and the mean ± s.e.m. is shown.

Vision things

Philip Campbell

Science's unpredictability has not prevented a group of invited scientists from being farsighted about future possibilities in fundamental research and its applications.

Anticipation is one thing, vision quite another. Geneticists and others are relishing the prospect of the maps and inventories that are to come, and the inevitable insights into organismal development and function, relationships between species and between kingdoms, and the evolutionary past. But where's the new vision? And what sorts of visions are driving other parts of biology and other sciences towards new discoveries and technologies?

This *Nature* Insight provides a partial answer to those questions. The editors of *Nature* and of the ten *Nature* journals (including three new review journals), who usually thrive on mutual independence, have collaborated to select ten topics to help all of our readers peer into the future. In December 1999 such a survey anticipated the impacts, on research and on society, of foreseeable science. This time we look towards the unforeseeable, by focusing on what might take us there: cutting-edge basic science that might lead to unexpected technologies, and adventurous technologies that should lead to unpredictable, fundamental discoveries. Our authors have been given freedom to speculate, untrammelled by peer review. These, then, are personal visions, and can be considered to be somewhat courageous. After all, most readers should be around long enough to evaluate their prescience.

Nurturing foresight

The inherent dangers of scientific prediction were discussed in the previous survey¹. But visions are surely to be encouraged. Perhaps the most forward-looking vision ever was that of J. D. Bernal in 1929 in his book *The World, the Flesh and the Devil: An Inquiry into the Three Enemies of the Rational Soul*. He anticipated life transforming itself into an ethereal intelligence dispersed between the stars. The physicist Freeman Dyson demonstrated² subsequently that this was not altogether incredible within known physical laws. This collection of articles falls short of that sort of reach. But it does include two aspects of mind: the scientific connection between life and artificial intelligence (see Rodney Brooks, page 409), and future technologies based on the ability to control mechanical objects using neural prosthetics (Miguel Nicolelis, page 403).

Physics' legendary successes have been based partly on its ability to establish principles and then develop predictions and elaborations. Biology, in contrast, is currently thriving on its cataloguing of the particular. Seeking principles is fine in principle but is way down the agenda of most molecular and cell biologists. But as our knowledge of the molecular circuitry of cells grows, so grows the opportunity to organize that knowledge conceptually and to depict and hypothesize about larger organizational patterns and even — possibly — principles. What is the most we might aspire to in cell biology? On page 387, Jack Szostak and colleagues imagine in some detail what it will take to synthesize a whole cell. On page 391, Drew Endy and Roger Brent elucidate the scale of the challenges to come in computational simulation of intracellular networks and even entire cells.

A place for chemistry

Molecular and cell biology also provide opportunities for chemists, whose insights will be ever more necessary in understanding the molecular dynamics of cells. But despite rumours to the contrary, chemistry itself still provides opportunities for chemists too. It is particularly appropriate now, 100 years after quantum mechanics was conceived, to focus on quantum chemistry. Over the past decade or so, chemists have been exploring not only the fundamentals of chemical reactions but also subtle ways of steering them by using photons to select quantum states as reactions proceed. Stuart Rice anticipates the potential technological applications of such fundamental chemistry on page 422.

Particle physics stands at a critical juncture. Just when the Large Electron-Positron Collider at CERN, in Geneva, spotted what might have been a few Higgs bosons — the long-sought particles that give other particles the property of mass — it was closed down to make way for the next generation of accelerator, the Large Hadron Collider. Fermilab's Tevatron near Chicago may establish the Higgs particle's existence conclusively in the next few years, but theorists also believe we are on the threshold of a world of new fundamental laws. To probe them will require new accelerator technologies, which may themselves have other technological spin-offs, as have past accelerators. One such

scheme, at least two accelerator generations ahead, is described by John Ellis and Ian Wilson on page 431.

Star gazing

Like the high-energy physicists, astronomers too can look forward to technologies that will not only expand their horizons but also reveal new ones. Roger Angel is a highly creative developer of new telescopes for optical and infrared wavelengths. As he explains on page 427, technologies for large ground-based and space-based telescopes will extend enormously our capacity for unimagined discoveries, as well as our deeper appreciation of objects already identified.

In contrast, firmly embedded in the everyday world, there is a quite different sort of vision: that of technologies created around principles of structure and pattern established by nature itself. As Philip Ball describes on page 413, "proteins and elephants" can stimulate engineers of the future.

Wherever there is scientific vision, there is a danger of hubris — or at least perceived hubris. Dyson's cosmically eschatological visions of intelligence's future caused him to be criticized by philosophers for, as they perceived it, wanting to control the Universe. One vision that positively reeks of hubris is that of controlling our planet's surface conditions, on a global scale, to cope with the climatic and environmental changes arising from human activities. The possibilities deserve to be elucidated, and the concerns expressed. Enter Stephen Schneider, with David Keith, on page 417.

Another application of science, and one that has given rise to more than its fair share of social concern, is the genetic modification of crops. On page 397, George Blackburn views functional foods through the rather different lens of the physiology of under-nourishment.

It is surely one of scientists' roles to look unflinchingly at feasible extrapolations of current knowledge, however unpopular that may make them. If in so doing they stimulate our readers — including authors' communities as well as funding agencies — to lift their sights as well, so much the better. *Philip Campbell is the Editor of Nature and Editor in Chief of Nature publications.*

1. Campbell, P. *Nature* **402**(Suppl.), C7–C9 (1999).

2. Dyson, F. *Infinite in All Directions* (Harper & Row, New York, 1988).

Synthesizing life

Jack W. Szostak, David P. Bartel & P. Luigi Luisi

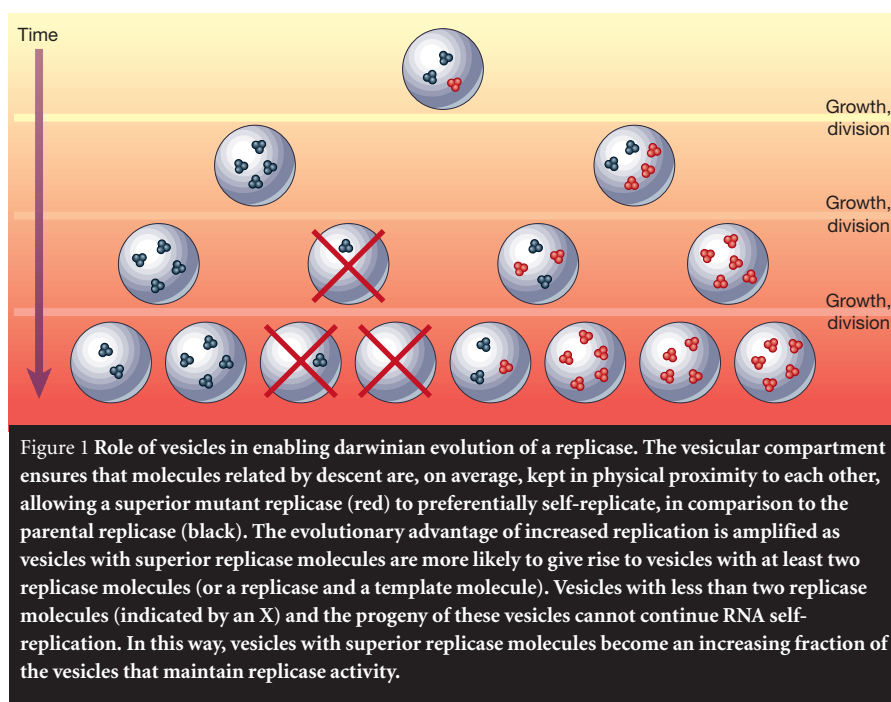
Advances in directed evolution and membrane biophysics make the synthesis of simple living cells, if not yet foreseeable reality, an imaginable goal. Overcoming the many scientific challenges along the way will deepen our understanding of the essence of cellular life and its origin on Earth.

The first challenge on the path to a synthetic life form is to imagine a collection of molecules that is simple enough to form by self-assembly, yet sufficiently complex to take on the essential properties of a living organism. Any 'stripping-down' of a present-day bacterium to its minimum essential components still leaves hundreds of genes and thousands of different proteins and other molecules¹. We must look to simpler systems if we hope either to synthesize a cell *de novo* or understand the origin of life on Earth. The search for simpler forms of life led to the 'RNA world' hypothesis² in which primordial cells lacking protein synthesis use RNA both as the repository of 'genetic' information and as enzymes that catalyse metabolism³. We believe that within this framework structures can be found that are both indisputably alive and yet simple enough to be amenable to total synthesis. We note that solutions found in the laboratory need not be chemically similar or even directly relevant to the actual molecular assemblies that led to the origin of life on Earth.

How simple can a cell be and still be considered as living? The answer depends on what we consider to be the essential properties of life. Defining life is notoriously difficult⁴; its very diversity resists the confines of any compact definition. An operational approach focuses on identifying simple cellular systems that are both autonomously replicating and subject to darwinian evolution. Autonomous replication is understood as continued growth and division which is reliant on the input of small molecules and energy only, and does not depend on the products of pre-existing living systems such as protein enzymes. Darwinian evolution requires the essential biological aspects of genetic variation and its phenotypic expression as variation in survival and reproduction.

Designing the protocell

We can consider life as a property that emerges from the union of two fundamentally different kinds of replicating systems: the informational genome and the three-dimensional structure in which it resides. The simplest way to enable darwinian evolution is to begin with a nucleic acid genome.



Although there is considerable debate about the nature of the first genetic polymers⁵, there is no doubt that RNA and DNA are the only currently practicable genetic materials for directed evolution *in vitro*. A growing body of experimental work points to the feasibility of evolving and/or designing in the laboratory an RNA replicase — an RNA molecule that can act both as a template for the storage and transmission of genetic information, and as an RNA polymerase that can replicate its own sequence^{6–8}. This molecule will be one of the key components of any synthetic cell.

But a replicase molecule by itself is not living, for two reasons. First, a single molecule could not actually replicate, as it cannot be both template and polymerase at the same time. Replication requires two RNA molecules — a replicase that acts as the polymerase, and another molecule, which could be either an unfolded replicase or an RNA complementary in sequence to the replicase, to act as a template. Some form of physical compartmentation is therefore required to keep replicase and template together, unless both are present at high concentration. Second and more subtly, a population of

replicases free in solution could not evolve into more active or accurate replicases. In solution, better replicases would replicate other RNA molecules more efficiently, but would have no advantage themselves, and would not increase in relative abundance. Again, some form of compartmentation is required: by keeping molecules that are closely related together, advantageous mutations can lead to preferential replication. After a period of replication, mutation and random assortment, some compartments will be occupied by mutant replicases, and others by the original replicases; better replicases will therefore be able to replicate each other more efficiently, giving them an overall advantage (Fig. 1).

All known cells use membranes composed of amphipathic lipids as their compartment-defining barriers, and therefore the easiest way to construct our simple protocell is to surround it with a lipid membrane. This also makes it easier to imagine how a simple cell could evolve into more complex cells, similar to present-day cells, without major architectural transitions. However, in the absence of any machinery for

cell growth and division, the membrane-bounded vesicle must itself be a spontaneously replicating entity.

So our simple protocell will consist of an RNA replicase replicating inside a replicating membrane vesicle. Both these components are self-assembling; the catalytically active structure of the replicase will form spontaneously as a consequence of its nucleotide sequence, while membrane vesicles assemble spontaneously as a result of interactions between the lipid molecules. As RNA molecules can become spontaneously encapsulated in vesicles as they form, the protocell as a whole could self-assemble. With compartmentation, the replicase component is not only capable of, but also inevitably subject to, variation, natural selection and thus darwinian evolution.

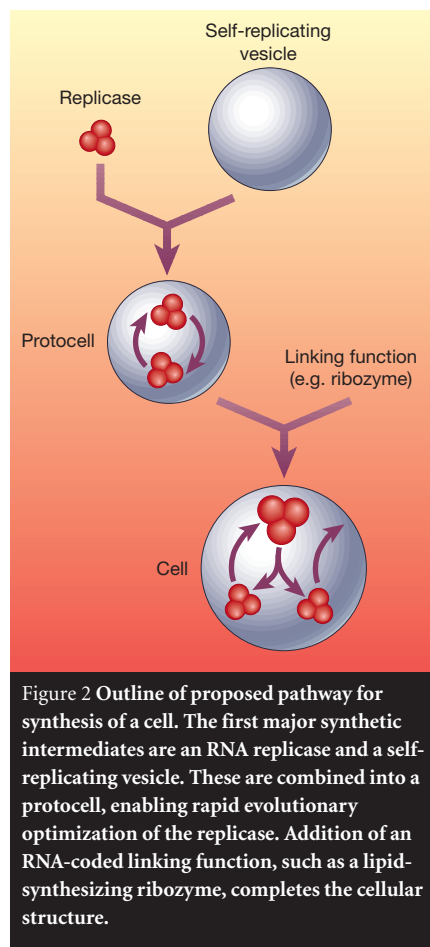
From protocell to living cell

Such simple protocells would be nearly, but not quite, alive. When fed small-molecule precursors for membrane and RNA synthesis, they would grow and divide, and improved replicases would evolve. However, a vesicle carrying an improved replicase would itself not have improved capacity for survival or reproduction. For this to happen, an RNA-coded activity is needed that imparts an advantage in survival, growth or replication for the membrane component. A simple example would be a ribozyme that synthesizes amphipathic lipids and so enables the membrane to grow. The membrane and the genome would then be coupled, and the 'organism' as a whole could evolve (Fig. 2) as vesicles with improved ribozymes would have a growth and replication advantage. A simple cell with an interdependent genome and membrane would be a sustainable, autonomously replicating system, capable of darwinian evolution. It would be truly alive.

The RNA replicase

The first experimental challenge is the evolution or design of an RNA replicase. Early attempts to derive an RNA replicase from the natural group I self-splicing introns produced ribozymes that could direct the assembly of oligonucleotide substrates on a template, and even the assembly of full-length RNA strands complementary to the ribozyme itself^{9,10}. The low efficiency of the reaction, however, even when driven by vast substrate excess, suggested that it was probably essential to use activated nucleotides, such as nucleoside triphosphates, to provide an energetic driving force for polymerization. The use of oligonucleotide substrates would also make it difficult or impossible to maintain a high concentration of all the different substrates needed for the replicase to mutate and evolve.

No natural ribozymes are known that can catalyse the required chemistry and use nucleoside triphosphates as substrates. As this is a complex enzymatic function, attempts to



evolve an RNA polymerase ribozyme experimentally have proceeded incrementally. First, *in vitro* selection was used to isolate, from a fully random set of starting sequences, a set of ribozymes that would carry out the correct chemistry—the attack of a 3'-hydroxyl on the α -phosphate of a triphosphate, yielding a new phosphodiester bond—but in a more favourable context, the joining, or ligation, of two RNA sequences. This experiment yielded many ligases that carried out pyrophosphate-activated oligonucleotide-ligation reactions¹¹. Of these, the class I ligase ribozyme synthesized the desired 3',5'-phosphodiester linkage in a template-directed reaction, and therefore carried out the same chemical transformation as protein-enzyme polymerases¹². Mutations that increased catalytic activity¹³ gave a highly active ligase that could join oligonucleotides with a rate of greater than one joining event per second^{12,14}. Derivatives of this ribozyme were subsequently shown to act as primitive polymerases capable of template-directed extension of a 'primer' strand of RNA complexed to the RNA template, using nucleoside triphosphates as substrates¹⁵. In essence, the oligonucleotide providing the 5'-triphosphate could be replaced by a single nucleotide, which can still be ligated to the 3'-end of the growing primer. The cycle of primer extension can be repeated several times before steric constraints prevent

further chain growth. These advances bring the evolution of a true RNA replicase tantalizingly close.

What further improvements are required to obtain an RNA replicase suitable for incorporation into an artificial cell? In its current form, the polymerase ribozyme recognizes the primer-template complex through hybridization to a particular unpaired segment of the template. This pairing restricts ribozyme movement along the template and reduces severely the number of compatible template sequences. A ribozyme that can recognize the primer-template using non-sequence-specific contacts would enable more extensive and general RNA synthesis. The next hurdle will be to improve the fidelity and efficiency of polymerization. It is possible that only a 100-fold increase in the rate of polymerization and a 10-fold improvement in the Watson-Crick fidelity of this ribozyme would lead to an RNA polymerase able to faithfully copy templates of its own length⁷.

A more subtle problem is that a true replicase must function both as a polymerase and as a template. How could the same RNA sequence act both as a highly active ribozyme structure, presumably favoured by stable folding, and as a template available for copying, favoured by less stable folding? A potential solution comes from the finding that active ribozymes can be reconstituted by the spontaneous self-assembly of two or more oligonucleotides; the separate oligonucleotides can be more or less unstructured, while the assembled complex can be stable and enzymatically active¹⁶. This solution has the advantage that the average length of sequence that needs to be copied by the replicase can be fairly short (30–40 nucleotides), but the potential disadvantage that the relatively unstructured + and – strand fragments might rapidly reanneal to form double-stranded RNA.

This leads to the issue of strand separation during or after replication in order to reconstitute the replicase. In principle, thermal denaturation might separate the strands of a double-stranded replication product. The extreme stability of long RNA duplexes, especially in the presence of significant concentrations of divalent cations, makes this approach unattractive, however, as conditions that would lead to denaturation would probably also lead to chemical degradation of the RNA and disruption of the membrane vesicle. But the alternatives involve more complexity in the RNA replication machinery—either RNA helicase activities to carry out energy-dependent strand separation, or, as is the case with bacteriophage T7 RNA polymerase, a portion of the replicase that binds single-stranded RNA and peels the new RNA strand off the template as it is synthesized. The rapid formation of local secondary structure in the RNA strands would then prevent them re-forming a dead-end full-length duplex.

The RNA polymerase (a protein enzyme) of the Q β virus is known to depend on the secondary structure of the viral + and – strands to kinetically block duplex formation¹⁷. A better understanding of the kinetics of RNA folding will be required to predict sequences that will be reasonably stable as unpaired + and – strands when packaged in the same vesicle.

An attractive alternative strategy is replication by strand-displacement on a duplex template, which provides a mechanism for coupling the chemical energy released during polymerization to strand displacement. Replication would also be more likely to start at the beginning of the genome, because of transient ‘fraying’ of the termini of the duplex. However, specific oligonucleotide primers may be required to obtain significant initiation. These could be difficult to deliver to the vesicle interior and might not be considered ‘small-molecule’ substrates. Nevertheless, the use of primers would allow replication of the entire template, avoiding gradual shortening without having to enlist a telomerase-like activity or having to prime polymerization with a single nucleotide at the 3′-terminus of the template.

The class I ligase is an excellent starting point for attempts to evolve a replicase but does have one drawback. Its minimal catalytic domain is about 100 nucleotides long. When coupled to additional domains that may be required for proper template binding, fidelity and strand separation, the total length could approach 200–300 nucleotides. The longer the replicase, the more difficult the problem of replication, so shorter replicases should be looked for. One approach would be to use more highly activated nucleotide substrates, such as the phosphorimidazolides. As these substrates are more reactive, less rate enhancement would be required to achieve extension rates in the range of one nucleotide per minute or per second, and a simpler and shorter ribozyme might suffice. Clearly, there are many possible approaches to evolving an RNA replicase, all of which bear investigation.

The membrane compartment

The vesicle component of a protocell or simple cell must possess a suite of rather unusual properties, including spontaneous growth, spontaneous division, permeability to nucleotide substrates, physical stability under conditions required for RNA replication and compatibility with ribozyme activity.

Spontaneous vesicle growth could in principle occur either gradually by the incorporation of single lipid molecules or micelles, or stepwise by fusion with other vesicles. Gradual growth, which is more biological, is possible if the rate of incorporation of lipid into pre-existing vesicles is greater than the rate of spontaneous assembly into new vesicles. Alternatively, the catalytic gen-

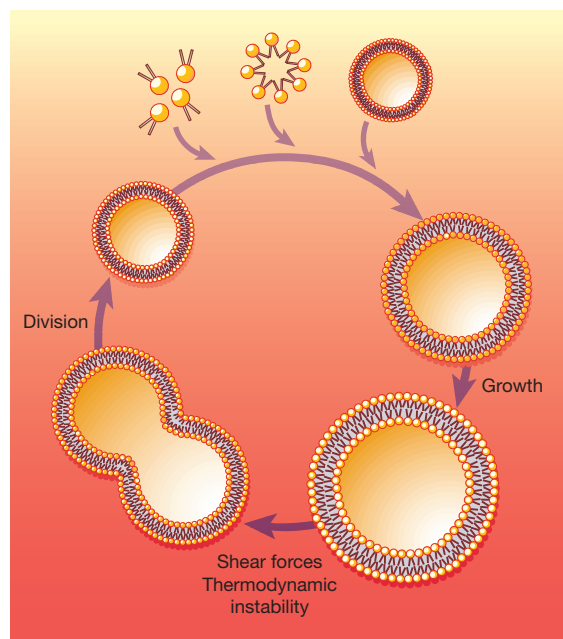


Figure 3 Modes of vesicle growth and division. Self-replicating membrane vesicles can grow either gradually or in discrete steps, and may divide either spontaneously or under the influence of external environmental forces.

eration of new lipid molecules in a vesicle membrane could also lead to vesicle growth. Fatty acids such as oleate form micelles at high pH, an oil phase at low pH, and bilayer membrane vesicles at intermediate pH — these are thought to be stabilized by hydrogen bonding between the protonated and ionized carboxylates of the fatty acid. When the pH of a preparation of oleate micelles is lowered to the pK of oleate in a membrane (pH ~8) the micelles slowly rearrange to form vesicles. However, pre-existing oleate vesicles greatly accelerate the rate of formation of new vesicles¹⁸. Remarkably, the newly formed vesicles have a size distribution similar to that of the preformed vesicles, which can be quite different from the size distribution in the absence of preformed vesicles. The mechanism of this ‘matrix’ effect remains unknown, although some form of growth and division is an intriguing possibility. Subsequent work showed that oleate vesicles seem to grow in size in the presence of oleic anhydride, presumably as a result of vesicle-mediated catalysis of hydrolysis of the oleate precursor, followed by incorporation of the newly generated oleate into the vesicle membrane¹⁹.

These experiments could not distinguish between pre-existing vesicles that grew by incorporating new lipid, and newly formed vesicles. More recently, however, preformed oleate vesicles have been tagged by preparation in the presence of ferritin²⁰. After exposure of these vesicles to additional oleate micelles, electron microscopic examination revealed that the tagged vesicles had grown to a larger average size, providing strong evidence for vesicle growth by spontaneous incorporation of new lipid.

Vesicle growth in discrete steps can occur by vesicle-vesicle fusion. The fusion of vesicles composed of acidic phospholipids is

mediated by low concentrations of Ca²⁺ ions^{21–23}, and short fusogenic peptides can catalyse the fusion of vesicles composed of neutral phospholipids²⁴. The instability of small strained vesicles can provide a thermodynamic driving force for this process. However, these processes tend to operate most efficiently under conditions that are very close to those favouring a phase change for the component lipids, and thus catastrophic loss of vesicle integrity. The narrow range of conditions under which fusion is efficient but vesicle disruption is minimal indicates that it may be difficult to devise a robust cycle of growth and division using this approach. On the other hand, vesicle fusion could bring fresh supplies of nucleotide substrates to replicases encapsulated in a separate vesicle.

What about division? In the absence of the complex machinery that controls the division of modern cells, the division of growing vesicles must rely on the intrinsic properties of the vesicle and the physical properties of the environment²⁵. Input of energy from the environment can generate a population of vesicles with a non-equilibrium size distribution. High environmental shear forces, for example, can cause vesicles to divide. Such a process, operating in conjunction with a spontaneous growth mechanism, could lead to a primitive cell cycle controlled entirely by the biophysical properties of the membrane and environmental forces. An intriguing possibility is that the process of division could be highly favoured, or even become spontaneous, with lipid compositions that yield vesicles of optimum size for thermodynamic stability (Fig. 3).

Addition of lipid to lipid micelles causes micellar reproduction because the growing micelles become thermodynamically unstable above a certain size^{26,27}. For example, ethyl caprylate oil layered over alkaline water hydrolyses slowly at the interface. Once the

critical micelle concentration for caprylate is reached, substrate is solubilized in the micelles and the rate of hydrolysis greatly increases as a result of the increased area of lipid–water interface, resulting in a rapid increase in the number of micelles. Whether an analogous system for the spontaneous growth and division of vesicles can be achieved remains an open question.

The membrane of a synthetic cell should allow transport of small-molecule substrates such as nucleotides, while keeping macromolecules encapsulated. The demonstration of RNA synthesis catalysed by polynucleotide phosphorylase inside vesicles, using external nucleoside diphosphate substrates, indicates that this requirement may not be too stringent^{28,29}. Short-chain lipids, lipid mixtures, and co-surfactants such as cholate can all make membranes more permeable to small molecules^{30,31}. Amplification by polymerase chain reaction of DNA inside vesicles suggests that it may be possible to find vesicle compositions compatible with conditions for RNA replication³².

An alternative approach to feeding the replicase small-molecule substrates would be to encapsulate the substrates within vesicles, which could then be delivered to the replicase by vesicle fusion. With repeated cycles of fusion and fission, a replicase initially present in just a few vesicles would spread throughout the vesicle population. Although the vesicle component would not be growing and evolving, the replicase component would be. In some ways, the replicase of this system is analogous to bacteriophage, which clearly evolves as it propagates through a system of compartments (bacterial cells). Such a system could provide a powerful tool for the evolution of replicase activities.

Replicase-vesicle coupling

Once robust self-replicating replicases and vesicles have been devised, they must be brought together in a compatible and interdependent union in which the vesicle–RNA system as a whole is subject to darwinian evolution. First, the replicase and vesicle must be compatible; the replicase must be able to replicate inside the vesicle, and the vesicle must be able to grow and divide unperturbed by its cargo of RNA. It is impossible to foresee all the problems, but some are already clear. Most obviously, both replication cycles must operate under a single consistent set of conditions. Many ribozymes have optimal activity in the presence of high concentrations of divalent metal ions. In such conditions, vesicles composed of acidic phospholipids would aggregate, possibly interfering with growth and division. The timescales of the RNA and vesicle replication cycles must also be approximately the same, so that the replicases can keep up with the vesicles but not reach such a high internal concentration that vesicle growth or division is disrupted.

Once a compatible membrane and replicase have been united to form a protocell, an additional RNA activity must emerge spontaneously or be added exogenously to couple the information coded by the genome to the fitness of the vesicle. The emergence of ribozymes that catalyse the synthesis of suitable amphipathic molecules would facilitate growth of the membrane compartment, and they could ultimately be used to alter and control the properties of the vesicle membrane. Structural RNAs with favourable interactions with the interior vesicle wall could stabilize the vesicle and lead to preferential survival. The evolution of RNA filaments analogous to actin filaments or microtubules could influence vesicle shape and dynamics, and begin to provide internal control of cell division. It is clear that genomic (that is, RNA) variation would influence the ability of the cell to grow, survive and reproduce, and thus would control the fitness of the cell as a whole.

With coupling accomplished, the living synthetic cell would be capable of evolving in ways that none of its components are, and we expect that the strong selective forces to which it will be subject will provide a powerful driving force for the emergence of biochemical complexity, which in turn will lead to increasingly tight coupling and interdependence of the RNA and membrane components. The emergence of better replicases would allow the replication of longer RNA genomes, and the incorporation (or internal generation) of new random sequences would provide a source of new RNA activities. The evolution of ribozymes that contribute to the synthesis of RNA precursors would enhance the efficiency of the RNA replication process, both by decreasing the need for externally supplied substrates, and perhaps also by decreasing the need to spontaneously transport complete nucleotides across the membrane. Particular RNA sequences might also selectively alter the membrane permeability³³, increasing the intracellular availability of replicase substrates. Other ribozymes might lead to the formation of RNA–lipid conjugates, providing a mechanism for membrane anchoring of ribozymes and possibly more equal segregation of ribozymes into daughter cells. As the number of ribozymes and structural RNAs grows, the membrane compartment will be crucial in maintaining the spatial integrity of the assembly of cooperating RNA species.

Experimentally, the potential exists to jump-start the emergence of biochemical complexity by the *in vitro* selection and directed evolution of potentially adaptive ribozymes. Alternatively, by supplying a population of cells with random RNA sequences, one might observe the process of evolving complexity in real time, and thus determine experimentally what new ribozyme activities were most accessible and

advantageous for evolving simple cells. In the long run, it might even be possible to observe at least some aspects of the evolution of protein synthesis, possibly with different basis sets of amino acids. These experimental possibilities could provide fascinating insights into what is now a complete black box of early evolution.

Jack W. Szostak is at the Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. David P. Bartel is at the Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. P. Luigi Luisi is at ETH-Zentrum, Institut für Polymere, Universitätsstrasse 6, CH-8092 Zürich, Switzerland.

- Hutchison, C. A. *et al.* *Science* **286**, 2165–2169 (1999).
- Gilbert, W. *Nature* **319**, 618 (1986).
- Joyce, G. F. & Orgel, L. E. in *The RNA World* (eds Gesteland, R. E., Cech, T. R. & Atkins, J. F.) 49–77 (Cold Spring Harbor Laboratory Press, New York, 1999).
- Luisi, P. L. *Origins Life Evol. Biosphere* **28**, 613–622 (1998).
- Joyce, G. F. *New Biol.* **3**, 399–407 (1991).
- Hager, A. J., Pollard, J. D. & Szostak, J. W. *Chem. Biol.* **3**, 717–725 (1996).
- Bartel, D. P. in *The RNA World* (eds Gesteland, R. E., Cech, T. R. & Atkins, J. F.) 143–162 (Cold Spring Harbor Laboratory Press, New York, 1999).
- Bartel, D. P. & Unrau, P. J. *Trends Cell Biol.* **9**, M9–M13 (1999).
- Doudna, J. A. & Szostak, J. W. *Nature* **339**, 519–522 (1989).
- Green, R. & Szostak, J. W. *Science* **258**, 1910–1915 (1992).
- Bartel, D. P. & Szostak, J. W. *Science* **261**, 1411–1418 (1993).
- Eklund, E. H., Szostak, J. W. & Bartel, D. P. *Science* **269**, 364–370 (1995).
- Eklund, E. H. & Bartel, D. P. *Nucleic Acids Res.* **23**, 3231–3238 (1995).
- Bergman, N. H., Johnston, W. K. & Bartel, D. P. *Biochemistry* **39**, 3115–3123 (2000).
- Eklund, E. H. & Bartel, D. P. *Nature* **382**, 373–376 (1996).
- Doudna, J. A., Couture, S. & Szostak, J. W. *Science* **251**, 1605–1608 (1991).
- Priano, C., Kramer, F. R. & Mills, D. R. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 321–330 (1987).
- Blochiger, E., Blocher, M., Walde, P. & Luisi, P. L. *J. Phys. Chem.* **102**, 10383–10390 (1998).
- Walde, P., Wick, R., Frezza, M., Mangone, A. & Luisi, P. L. *J. Am. Chem. Soc.* **116**, 11649–11654 (1994).
- Berclaz, N., Muller, M., Walde, P. & Luisi, P. L. *J. Phys. Chem. B* (in the press).
- Ingolia, T. D. & Koshland, D. E. Jr *J. Biol. Chem.* **253**, 3821–3829 (1978).
- Wilschut, J., Düzgünes, N., Fraley, R. & Papahadjopoulos, D. *Biochemistry* **19**, 6011–6021 (1980).
- Ellens, H., Bentz, J. & Szoka, F. C. *Biochemistry* **24**, 3099–3106 (1985).
- Parente, R. A., Nir, S. & Szoka, F. C. Jr *J. Biol. Chem.* **263**, 4724–4730 (1988).
- Norris, V. & Raine, D. J. *Origins Life Evol. Biosphere* **28**, 523–537 (1998).
- Bachmann, P. A., Luisi, P. L. & Lang, J. *Nature* **357**, 57–59 (1992).
- Luisi, P. L. *Adv. Chem. Phys.* **XCII**, 425–438 (1996).
- Chakrabarti, A. C., Breaker, R. R., Joyce, G. F. & Deamer, D. W. *J. Mol. Evol.* **39**, 555–559 (1994).
- Walde, P., Goto, A., Monnard, P.-A., Wessicken, M. & Luisi, P. L. *J. Am. Chem. Soc.* **116**, 7541–7547 (1994).
- Ueno, M. *Biochemistry* **28**, 5631–5634 (1989).
- Albalak, A., Zeidel, M. L., Zucker, S. D., Jackson, A. A. & Donovan, J. M. *Biochemistry* **35**, 7936–7945 (1996).
- Oberholzer, T., Albrizio, M. & Luisi, P. L. *Chem. Biol.* **2**, 677–682 (1995).
- Khvorovova, A., Kwak, Y. G., Tamkun, M., Majerfeld, I. & Yarus, M. *Proc. Natl Acad. Sci. USA* **96**, 10649–10654 (1999).

Acknowledgements. We thank our colleagues and the members of our laboratories for helpful discussions and comments on the manuscript. J.W.S. is an Investigator of the Howard Hughes Medical Institute. This work was supported by grants from NASA and the NIH to J.W.S., grants from NASA and the NIH to D.P.B. and grants from COST Supramolecular Chemistry, Project D 11-2 to P.L.L.

Modelling cellular behaviour

Drew Endy & Roger Brent

Representations of cellular processes that can be used to compute their future behaviour would be of general scientific and practical value. But past attempts to construct such representations have been disappointing. This is now changing. Increases in biological understanding combined with advances in computational methods and in computer power make it possible to foresee construction of useful and predictive simulations of cellular processes.

In molecular, cellular and developmental biology, compact and elegant theories of the sort familiar in physics are rare; rather, explanations of phenomena are typically couched in natural language narratives that describe the interactions of large numbers of distinct molecular entities. In this essay, we define *model* as any representation of a system. Models are usually made up of abstractions that are easier to manipulate than the actual system. We are concerned with models of cellular processes whose internal descriptions match the molecular mechanisms by which those processes act. In particular, we are interested in models that incorporate knowable quantities, including the number or amount of different entities (for example, proteins, transcripts or regulatory sites) and the rates at which these entities react, and in which the entities and reactions are governed by physical laws. We define *simulation* as a representation that embodies information contained in a model, and that provides access to the model by allowing computation of system behaviour.

To give an example of this usage, physicists routinely use computer simulations to access and elaborate predictions of the dominant theoretical framework in high-energy physics, the Standard Model. In the biological examples we will discuss here, the information that constitutes a model might be described in words or systems of equations, but the simulations that provide access to the models will run on computers.

The models used in molecular, cellular and developmental biology are typically heuristic. They arise alongside the process of experiment and are inseparable from it. Because they are based on experiments in which perturbation of system components has had observed effects, the models typically contain embedded knowledge of causality and of the passage of time. In such models, time progresses from one experimentally defined causal step to the next (Fig. 1a, b). In the simulations discussed here, time is absolute (Fig. 1c).

Too ambitious, too soon

Past efforts to model behaviour of molecular and cellular systems over absolute time

typically were qualitatively incomplete or oversimplified compared to available knowledge, and quantitatively incomplete in the sense that key numbers were unknown. For example, even thoughtful, carefully constructed models posited the control of embryonic and somatic cell proliferation by a single cyclin whose degradation controlled entry into mitosis¹ after the existence of different cyclins that controlled progression through different phases of the cell cycle was established². Models of circadian rhythms based on known molecular entities³ were immediately outgrown as new molecules (for example, Clock and Cycle) were discovered⁴. In general, such models did not result in predictions of phenomena that biologists perceived to be significant enough to warrant subsequent experimental effort.

An extreme example of the disjunction between model and experiment is the study of imagined networks of mutually activating and repressing genes (or gene products) — so-called 'genetic regulatory networks'. Early studies showed that relatively simple interactions among network members could give rise to surprisingly complicated behaviour (ref. 5 and Fig. 2). However, by the early 1970s it was becoming apparent that few if any living systems had complex regulatory networks of this type, and that living systems regulate their transitions from state to state in other ways (see below). Although research on these imagined networks continues to this day, most biologists are either unaware of the work or ignore it.

But despite this history, we can now contemplate models of molecular, cellular and developmental biological systems that are coupled to experiment and result in increased understanding. One reason for optimism is that for some processes, enough biology is now known to begin to constrain useful models, and we can foresee obtaining much of the rest.

Qualitative simulations

One computable representation we may shortly expect to see is the so-called Biological Information System (BIS). The term

comes by analogy to Geographical Information Systems (GISs). BISs will extend current databases by embodying largely qualitative mechanistic knowledge. Over the next decade, BISs are likely to develop further, to encompass all known qualitative facts, including the components, their interactions and causal relationships for entire cellular subsystems and cells. The qualitative relationships among the components of such systems would be described by natural language equivalents — a small group of verbs defining permitted interactions. The information contained in BISs will be used to compute qualitative system behaviour over small numbers of causal steps. Although such computations will be only simple manipulations of existing knowledge, they will still be useful (see below).

Quantitative models

Progress in computation

Quantitative models of cellular processes often involve the representation of chemical reactions for which reactant molecules are scarce, and the continuous-variation approximation of differential calculus breaks down. Whereas in the 1950s the advent of the digital computer allowed numerical solution of large systems of differential equations, it was not until the 1970s that stochastic methods⁶ were developed to handle scarce reactants. During the 1990s these methods began to be applied widely to simulate biological systems^{7,8}. Recently, the efficiency of these methods has been increased significantly⁹, so that the earlier simulations⁸ can now be solved on desktop machines instead of supercomputers. Further reductions in computational cost will come from 'linking' deterministic and stochastic regimes (ref. 10 and D. T. Gillespie, personal communication), and may come from new methods that better handle large numbers of coupled reactions.

The promise of these methods also depends on increases in computing power. For example, one can now use a Gibson-modified Gillespie algorithm⁹ to execute 10¹⁰ reaction events per day on an 800-MHz

Pentium III processor. That is, a day on a current personal computer is sufficient to simulate 100 minutes of a 100-reaction system in which each reaction occurs ~20,000 times per second. At such speeds, a hypothetical simulation of *Escherichia coli* that tracked 10^{14} – 10^{16} reaction events per cell doubling (R.B. and D.E., unpublished data) could now run on existing multiprocessor computers¹¹ and, perhaps, on the single processors of 2020 (ref. 12).

Need for more information

Before we can model cellular systems quantitatively, we first need to overcome gaps in understanding. Even qualitative understanding is incomplete. For example, 33% of phage λ proteins and 45% of phage T7 proteins remain uncharacterized (L. Thomason, personal communication), but many of these probably contribute (at least quantitatively) to system behaviour under some conditions. In higher organisms, for complex traits (for example, disease susceptibility in humans and crop yield in plants), more of the contributing proteins remain unidentified and causal relationships are less worked out.

Second, we need to understand better the physics of some intracellular phenomena. In *E. coli*, for example, intracellular protein concentration is 200–320 mg ml⁻¹ (ref. 13). As biologists have long been aware^{13,14}, these high concentrations may well affect permitted reactions and their rates. Consistent with this idea, measurements of the diffusion of the *Aequorea victoria* green fluorescent protein (GFP) in *E. coli* reveal an apparent diffusion coefficient that is 11 times lower in *E. coli* than in water¹⁵. The same experiments also revealed that tagging GFP with a His₆ moiety lowered its apparent diffusion coefficient by another 40%, demonstrating that the behaviour of this widely used protein in the cellular environment is not well understood.

Third, even for biological systems in which all the components are known, we seldom understand precisely how they interact to make the process work. For example, the scaffold–kinase complex in one of the yeast signal transduction pathways is commonly thought of as a single protein complex. But consider that the scaffold protein (Ste5) can itself form at least a dimer, each monomer of which could form individual complexes with three different kinases, which between them have at least seven phosphorylation sites. These facts indicate that there can be 25,666 unique species that contain Ste5. Not only is this more than the number of Ste5 molecules in the cell (~5,000; ref. 16), but in principle each different species could have different quantitative or even qualitative functions. A similar problem occurs with combinations of regulatory proteins bound to complex

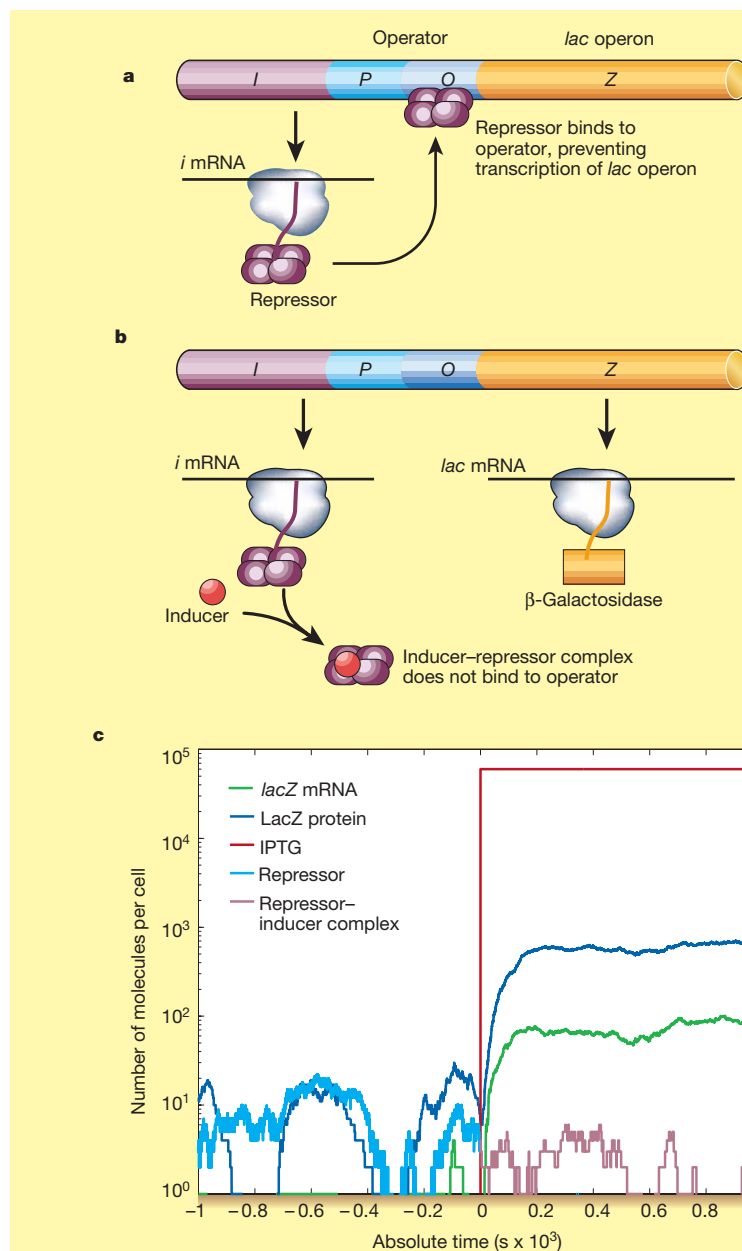


Figure 1 The *lac* operon of *Escherichia coli*. **a**, Qualitative model showing that without inducer, the *lac* repressor binds to the operator and LacZ is repressed. **b**, Qualitative model showing that in the presence of inducer, the *lac* repressor does not bind to the operator, *lac* operon mRNA is transcribed, and LacZ is produced (Y and A proteins are not shown). **c**, Output of a single run of a stochastic simulation based on this model. Note the fluctuation in number of *lac* repressor molecules, which is due to variation in timing of individual synthesis and decay events. At time zero, the concentration of inducer (IPTG) rises to 4×10^{-4} M; *lac* repressor is inactivated within seconds and induced transcription begins, followed soon after by translation of *lac* mRNA and the appearance of the first new LacZ proteins.

gene-regulatory regions often found in eukaryotes. Defining which species are important, determining their qualitative and quantitative function, and keeping track of them in models (ref. 17 and L. Lok, unpublished results) all present considerable observational and computational challenges.

Fourth, we need to learn to make better use of physics to constrain models, and to

define key experiments. At present, perhaps the best examples come from studies of bacterial chemotaxis¹⁸. Consideration of process physics has also aided the understanding of other sensory systems. For example, animal visual systems should ultimately be limited in sensitivity by the rate at which the retinal moiety of rhodopsin isomerizes as a result of thermal noise, and experiments in fact reveal that cold toads actually see better in dim light

than warm ones¹⁹. Other applications of physics are less obvious. For instance, particular models of circadian rhythms are sensitive to noise and, given variation in the timing of reaction events, do not produce the stable oscillations observed in nature²⁰. Thus, for circadian rhythms, consideration of system function and likely process physics helped to dismiss a particular class of models. Looking beyond these cases, it is interesting to note that much of cell and organismic biology can be understood as the processing of information from the genome, from internal events, and from external events, by an amorphous 'architecture' of diffusing molecular components. We can thus hope that future developments in information theory will provide broader insights into biological function and help constrain models and suggest experiments.

Finally and most importantly, we need to devise new experimental methods for obtaining quantitative data about biological processes. At the moment, we lack good experimental means to determine: (1) the absolute numbers of different molecular species in populations of cells; (2) the numbers of these species in individual cells; (3) how those numbers vary among individual, genetically identical members of a population; (4) how those numbers vary over time; and (5) the rates of the individual reactions causing that variation. The development of methods for acquiring this quantitative knowledge is one of the greatest challenges for biology in the twenty-first century²¹, one well beyond the scope of this essay.

Need to choose useful levels of resolution

Any model embodies a physical and logical level of resolution. It seems likely that for many cellular and early embryonic developmental processes, the appropriate level of resolution is that of known proteins, DNA regulatory sites, and so on. However, in any given instance, an assumption that those molecules are 'localized' to well-mixed compartments may not be sufficient. For example, the discovery that *E. coli* MinC and MinD, proteins that suppress septum formation and cytokinesis, are localized to the poles, helped explain why the cell normally divides in the middle^{22,23}. But the same experiments revealed the startling fact that individual protein molecules do not stay put. Molecules of MinC and MinD translocate from one pole to the other over tens of seconds. The use of fluorescent fusion proteins (and other methods) will surely reveal many instances where spatial localization is important for understanding process function. It is thus likely that future simulations will need to divide the cellular milieu into individual voxels (volume elements) in which reactions occur.

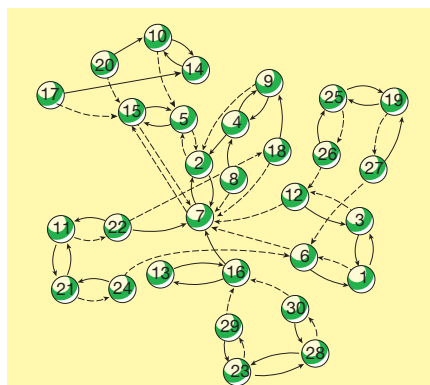


Figure 2 Behaviour of a 'genetic regulatory network'. In this example, regulation comes from cross-acting activators and repressors, and each of the several hundred genes is regulated by the products of two others. Shown are transitions among the 30 stable states of this network⁵. State transitions in such networks show, for example, basins of attraction and chaotic regimes⁴². In general, living systems seem to use other mechanisms to regulate their transitions from state to state (see text).

Future simulations will also need to allow for transition among different levels of resolution. A biologist might describe a protein as a simple ellipsoid, then, in the next breath, explain the effect of a point mutation by the atomic-level structural changes it causes in the active site. We can imagine a future simulation of an intracellular signalling pathway that ignored the shape and size of the individual molecular components, except when computing the effect of a kinase inhibitor when specific atomic information would be required about the interaction of the inhibitor with an active site. Similarly, future simulations of a cell (or groups of cells) might treat individual signal transduction pathways as parameterized modules, until pathway-specific effects needed to be represented. By allowing transitions to the coarsest level of resolution needed to represent observed behaviour, future simulations will use fewer computer cycles, and facilitate the ability of researchers to comprehend and interact with them.

Need to interact with experiment

Just as biological models were developed through the comparison of model-based predictions with experimental observations, so simulations of biological systems will need to develop alongside of, and in comparison with, experiments. The level of comparison will sometimes be qualitative. For example, a simulation of T7 growth²⁴ allowed the prediction that some rearranged genomes should encode phage that grow faster than wild type (a prediction that for the single genome tested to

date proved incorrect²⁵). At other times the level of comparison will be quantitative, based on static endpoints. For example, Kananyan *et al.*²⁶ and Arkin *et al.*⁸ compared the computed number of λ phage that form lysogens as a function of multiplicity of infection to experimental observations of Kourilsky *et al.*²⁷. In the future, the important level of comparison may frequently be quantitative, based on time-dependent behaviour. For example, discrepancies between computed and observed phage T7 protein synthesis rates suggested that translation from some T7 messenger RNAs might be subject to negative regulation by an as-yet-unknown mechanism (Fig. 3 and ref. 25).

There are very few biological systems for which complete quantitative models can be constructed from existing information. Because contemporary biologists have no shortage of hypotheses they find worth pursuing, virtually all generated without recourse to quantitative models, the information needed to construct them will not automatically be forthcoming. Thus, to be successful, future modelling efforts will probably need to direct and influence ongoing experiments.

Need to define model systems

Sometimes it will be easier to gather experimental data from simplified systems. Consider the previously mentioned yeast signal cascade. If there are 25,666 unique protein complexes that contain Ste5, and it is unknown which occur *in vivo*, and experimental determination of complex existence is not easy, we might reduce the number of unique complexes by fusing individual protein monomers into chimaeras that retain biological function (P. M. Pryciak, personal communication). Simplified systems can even be constructed from scratch. For example, several groups have constructed simple genetic systems using prokaryotic repressors. So far, the synthetic systems constructed have been relatively simple, with around a dozen genetic components^{28,29}. However, as the ability to synthesize and assemble large DNA fragments³⁰ continues to increase (and the cost of synthesis decreases), more ambitious systems will be designed and constructed.

Still, a good deal of information for quantitative models will be gathered from non-simplified systems — cells and organisms. The organisms and cell types may or may not be well studied. But in this genomic age, if it seems appropriate to develop a hitherto understudied organism into an experimental system, we can at least hope to bring about a basic level of understanding by marshalling the full power of sequencing, gene expression monitoring, large-scale mapping of protein interactions, functional

analysis by transposon mutagenesis, deletion mutagenesis, and dominant protein-based approaches²¹.

Control of behaviour by genes

Figure 4 shows an eighteenth-century orrery, a quantitative simulation of the motion of the planets in the Solar System. Although the observations on which the simulation was based and the understanding of the physical laws that governed its elements were, in retrospect, quite accurate, the computed positions of the planets eventually deviate from what is observed. Deviation from observation is due both to imperfections in the clockwork, the brass rings and gears, and to the fact that, over long periods of time, the motion of the planets around the Sun is chaotic³¹. Although perhaps not chaotic, in biological systems (and simulations), too much depends on chance interactions among small numbers of interacting molecules to yield behaviour that is completely determined over time.

However, aspiring modellers can make use of the fact that cells and organisms use a number of genetic mechanisms to supplement their highly imperfect biochemical clockwork and keep their dynamic behaviour on track. First, biological systems frequently go back to the genome, invoking subprogrammes that reset them into new states. Regulatory proteins, frequently gene activators, initiate these genomic subprogrammes. For example, expression of MyoD protein initiates a course of gene expression that converts fibroblasts into myoblasts (muscle precursors)³². Similarly, ectopic expression of the Eyeless protein in the future leg, wing or antenna tissues of developing *Drosophila melanogaster* larvae invokes a subprogramme that results in (nonfunctional) eyes at the sites of Eyeless expression³³. Once initiated, progression through any given process may rely on biochemical clockwork. But at some point the subprogramme is completed, and progression to the next process presumably requires invoking a new subprogramme.

Second, cells use checkpoint controls — feedback mechanisms that prevent a sequence of events from starting, and hold the cell at the 'checkpoint' until the mechanism receives a signal that a required sequence of earlier events has in fact been completed. Checkpoints are defined operationally, for example by mutations that arrest progression of cellular systems at given states. For example, in the yeast *Saccharomyces cerevisiae*, the checkpoint protein Rad9 prevents cells with DNA damage from attempting a new round of DNA synthesis until the damage is repaired³⁴. Both genomic subprogrammes and checkpoint controls punctuate the temporal transitions of systems and provide the opportunity to reset them to new starting states.

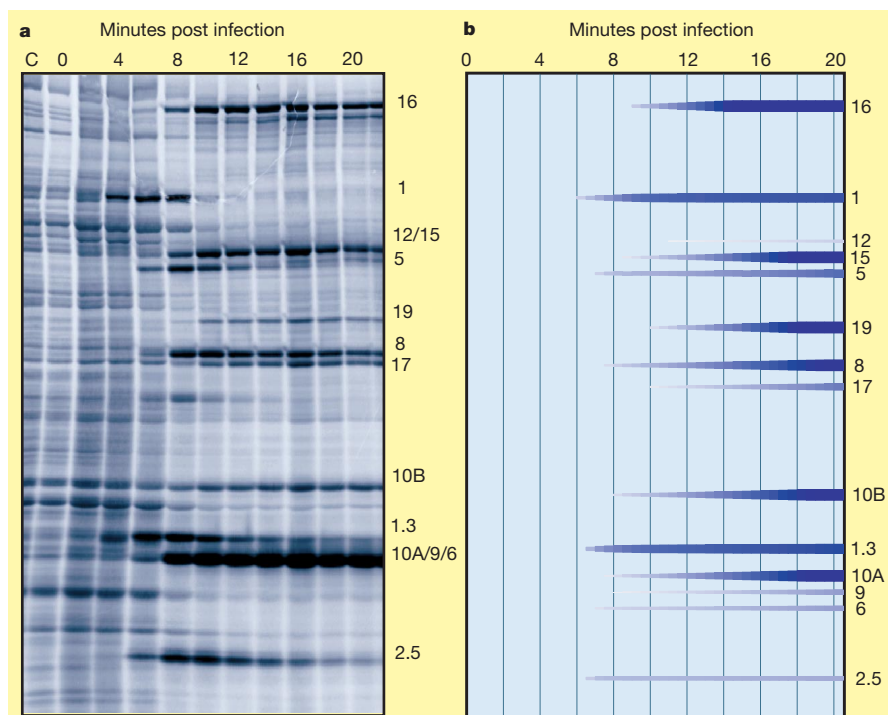


Figure 3 Observed and computed rates of phage T7 protein synthesis. **a**, Experimental determination of time and amount of phage protein synthesis during T7 infection (see <http://virus.molsci.org> for experiment and simulation details). **b**, Simulation output. The T7 simulation is based on current biological knowledge. Note that comparison of observed and computed T7 protein synthesis rates reveals that during an actual phage infection, synthesis of the T7 proteins gp1, gp1.3, gp2.5 and gp5 is negatively regulated by unknown mechanism(s).

Third, cells and organisms use other, less well understood mechanisms that coordinate timing of biological events and place dynamic system behaviour under more regulation than could be provided by biochemical clockwork alone. For example, *clk-1* mutants of the nematode *Caenorhabditis elegans* develop slowly at 15 °C, faster at 20 °C, and still faster at 25 °C. When two-cell *clk-1* embryos removed from 15 °C mothers are shifted to 20 °C, they continue to develop slowly, whereas two-cell embryos from 25 °C mothers shifted to 20 °C continue to develop rapidly^{35,36}. This observation shows that — beginning at a developmental stage before transcription of the embryo's own genes starts — the tempo of development in the wild-type worm is specified by a mechanism that is in part temperature independent. Construction of quantitative models can only further focus experimental attention on *clk-1* and other mechanisms that govern the timing of biological processes. Dedicated mutant hunts and 'protein genetic' screens³⁷ may reveal additional ways by which cells and organisms coordinate and regulate their time-dependent behaviour and reset to new states.

Consequences of success

The increasing amount of biological knowledge will probably itself be sufficient to force

the development of BISs to contain it. Such information systems will be good for more than teaching and learning. Relatively simple operations on (partly quantitative) information in GISs now allow people to determine driving directions and distance. By analogy, consider an information system that embodies known interactions and causal relationships among proteins that regulate cell division, and which could use that knowledge to enumerate those entities affected by perturbing the activity of different members of the protein network. Imagine a physician performing cancer therapy in 2020 who is looking at a listing of the changes in DNA sequence, gene expression and proteins in an individual tumour. The physician might use this information together with a BIS to support decisions on whether the inhibition of a particular protein kinase is likely to be useful for treating that particular tumour.

Vetting information

Another consequence would be an increase in the accuracy of biological information. This increase arises naturally from the fact that large-scale modelling efforts will require the combining of information from many different sources. Biology currently tests the validity of qualitative conclusions from different laboratories by mechanisms that range from peer-review to gossip. These are

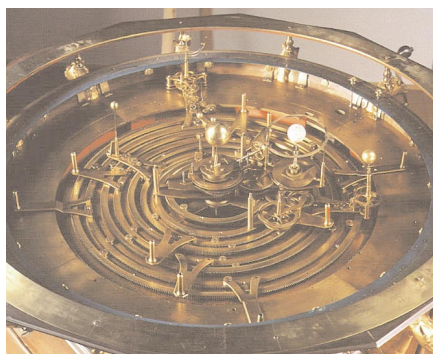


Figure 4 A simulation from 1773. Figure shows the workings of the 'Grand Orrery', a mechanical device that computes the positions of planets and moons in the Solar System (see <http://www.nmsi.ac.uk/collections/exhibits/george3/gallery.htm>). Less than 80 years later, models of planetary orbits were precise enough to demonstrate that deviations in the path of Uranus from its expected orbit could be accounted for by positing the existence of a new planet, and to tell astronomers where to point their telescopes to find it. Thus, by the 1840s, astronomical simulations were precise enough to allow prediction of Neptune. By contrast, during the 1990s no biological model of circadian rhythm allowed prediction of the regulatory proteins *Cycle* or *Clock*.

fairly effective; for example, they were able to demonstrate that not all 'phage T7 labs' were actually studying T7 (ref. 38). However, agreement on sets of quantitative information (and on very large sets of qualitative information) will probably require new ways of checking the accuracy, consistency and validity of that information. Making the computable information, the models and the simulations available to all scientists is clearly part of the solution. Once a draft simulation is constructed, discrepancies between computed and observed system behaviour will suggest changes to the model and new experiments. Done properly, providing access to simulations to large communities of biologists should accelerate the process of biological discovery itself.

Guiding intervention and therapy

Another consequence of success comes from the fact that quantitative mechanism-based models allow researchers to observe the complete behaviour of a specified system over time, and track all changes in its behaviour due to perturbations. It is easier to use a model to search for perturbations that have significant effects on system behaviour than it is to perform similar experiments on the living system. In some systems, an experimental search for sensitive components may not be possible. Moreover, models allow the search for multiple small perturbations that produce large effects when combined. In

most experimental systems, this is usually not possible.

Such capabilities will be useful for drug discovery and therapy. For example, quantitative models would help identify target proteins that give rise to therapeutic effects when partially inhibited. This alone would allow the development of small-molecule inhibitors that bind the target protein less tightly, thereby reducing the time needed to discover new drugs. Even more benefit may come from identification of cases where large changes in system behaviour could be achieved by partial inhibition of multiple protein targets. This would allow the identification of multiple targets that would permit the use of two or more drugs in smaller amounts, potentially resulting in fewer side effects. Models may also be useful in regimes (for example, anticancer therapy) in which drug concentration or amount of inhibition is limited. For example, models have been used to indicate that inhibition of a particular 'drug target', gene 1 messenger RNA of phage T7, has a paradoxical effect. The encoded protein, gp1, downregulates its own activity. Mutations in the messenger RNA that decrease 'drug' binding result in greater system inhibition³⁹.

Improving biological design

Models should form the basis of tools to aid in optimization of existing biological systems and design of new ones. Additionally, quantitative models will enable engineers to evolve biological systems by rounds of variation and selection for any function they desire. Such model-based evolution may complement existing organismic (for example, crossing two strains) and molecular (for example, mutagenesis using polymerase chain reaction, or DNA shuffling) approaches^{40,41} that depend on sometimes clever but often cumbersome selections and screens in the real world. As mentioned above, by the time computer-based optimization of living systems is possible, it will also be possible to fabricate large DNA sequences encoding the successful solutions, and thus to transfer successful designs from model to life.

Enabling new scientific understanding

Finally, mechanism-based models may bring now-unforeseen benefits to scientific understanding and capability. We have hinted at three of these. One comes from the fact that current biological understanding (and experimental methodology) does not deal very well with the passage of absolute time. The experiments needed to construct quantitative models, and consideration of those models, may help reveal mechanisms and insights into ways living systems regulate their temporal behaviour. A second comes from the idea that many biological systems can be described in terms of information processing. Quantitative models will be

needed to gain any insights from this metaphor. A third comes from the fact that mechanism-based models will be used as design tools and should speed the rise of a greatly heightened capability to engineer living systems. Although the lineaments of a world in which biology is directed by human intention might be foreseeable, the details of the changes to our selves and to our interaction with the living world cannot be foreseen.

Drew Endy and Roger Brent are at the Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA.

1. Tyson, J. J. *Proc. Natl Acad. Sci. USA* **88**, 7328–7332 (1991).
2. Lehner, C. F. & O'Farrell, P. H. *Cell* **61**, 535–547 (1990).
3. Leloup, J. C. & Goldbeter, A. *J. Biol. Rhythms* **13**, 70–87 (1998).
4. Rutala, J. E. *et al.* *Cell* **93**, 805–814. (1998).
5. Kauffman, S. A. *J. Theor. Biol.* **22**, 437–467 (1969).
6. Gillespie, D. T. *J. Comput. Phys.* **22**, 403–434 (1976).
7. McAdams, H. H. & Arkin, A. P. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
8. Arkin, A., Ross, J. & McAdams, H. H. *Genetics* **149**, 1633–1648 (1998).
9. Gibson, M. A. & Bruck, J. *J. Phys. Chem.* **2104**, 1876–1889 (2000).
10. Gillespie, D. T. *J. Chem. Phys.* **113**, 297–306 (2000).
11. van der Steen, A. J. & Dongara, J. J. <www.top500.org/ORSC/> (2000).
12. Hutcheson, G. D. & Hutcheson, J. D. *Sci. Am.* 54–62 (January 1996).
13. Cayley, S., Lewis, B. A., Guttman, H. J. & Record, M. T. *J. Mol. Biol.* **222**, 281–300 (1991).
14. Zimmerman, S. B. & Trach, S. O. *J. Mol. Biol.* **222**, 599–620 (1991).
15. Elowitz, M. B., Surette, M. G., Wolf, P.-E., Stock, J. B. & Leibler, S. *J. Bacteriol.* **181**, 197–203 (1999).
16. Bardwell, L., Cook, J. G., Chang, E. C., Cairns, B. R. & Thorner, J. *Mol. Cell. Biol.* **16**, 3637–3650 (1996).
17. Morton-Firth, C. J. & Bray, D. *J. Theor. Biol.* **192**, 117–128 (1998).
18. Block, S. M., Segall, J. E. & Berg, H. C. *Cell* **31**, 215–26 (1982).
19. Aho, A.-C., Donner, K., Hyden, C., Larsen, L. O. & Reuter, T. *Nature* **334**, 348–350 (1988).
20. Barkai, N. & Leibler, S. *Nature* **403**, 267–268 (2000).
21. Brent, R. *Cell* **100**, 169–183 (2000).
22. Hu, Z. & Lutkenhaus, J. *Mol. Microbiol.* **34**, 82–90 (1999).
23. Raskin, D. M. & de Boer, P. A. J. *Bacteriol.* **181**, 6419–6424 (1999).
24. Endy, D., Kong, D. & Yin, J. *Biotechnol. Bioeng.* **55**, 375–389 (1997).
25. Endy, D., Yu, L., Yin, J. & Molineaux, I. J. *Proc. Natl Acad. Sci. USA* **97**, 5375–5380 (2000).
26. Kananyan, G. Kh., Ratner, V. A. & Churavov, R. N. *Genetika* **16**, 2209–2017 (1980).
27. Kourilsky, P. *Mol. Gen. Genet.* **122**, 183–195 (1973).
28. Elowitz, M. B. & Leibler, S. *Nature* **403**, 335–338 (2000).
29. Gardner, T., Cantor, C. R. & Collins, J. J. *Nature* **403**, 339–342 (2000).
30. Yount, B., Curtis, K. M. & Baric, R. S. *J. Virol.* **74**, 16000–10611 (2000).
31. Sussman, G. J. & Wisdom, J. *Science* **257**, 56–62 (1992).
32. Davis, R. L., Weintraub, H. & Lassar, A. B. *Cell* **51**, 987–1000 (1987).
33. Halder, G., Callaerts, P. & Gehring, W. J. *Science* **267**, 1788–1792 (1995).
34. Weinart, T. A. & Hartwell, L. H. *Science* **241**, 317–322 (1988).
35. Wong, A., Boutis, P. & Hekimi, S. *Genetics* **139**, 1247–1259 (1995).
36. Branicky, R., Benard, C. & Hekimi, S. *BioEssays* **22**, 48–56 (2000).
37. Colman-Lerner, A. & Brent, R. *Trends Cell Biol.* 56–60 (Suppl. December 2000).
38. Studier, F. W. *Virology* **95**, 70–84 (1979).
39. Endy, D. & Yin, J. *Antimicrob. Agents Chemother.* **44**, 1097–1099 (2000).
40. Soong, et al. *Nature Genet.* **25**, 436–439 (2000).
41. Schmidt-Dannert, C., Umeno, D. & Arnold, F. H. *Nature Biotechnol.* **18**, 750–753 (2000).
42. Kauffman, S. A. *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford Univ. Press, 1993).

Acknowledgements. We thank R. Carlson, A. Colman-Lerner, D. Gillespie, M. Gruber, P. Pryciak, C. Kenyon, E. Kroll, E. Lyons, L. Lok, I. J. Molineux, O. Resnekov, T. Roosevelt, L. Thomason, J. Yin and L. You for useful comments, discussions or unpublished information. Work at TMSI is supported by grants to R.B. and D.E. from the NIH, DARPA and the Office of Naval Research.

Pasteur's Quadrant and malnutrition

George L. Blackburn

Basic science has been the wellspring of advances in technology in everything from cellular phones to nutrient-fortified foods. Biochemistry and food technology afford new techniques that enrich nutritional science and aid in the fight against malnutrition. Bionutrition, the term used to describe the application of these techniques, can be the source of innovations that will help eradicate malnutrition globally. Which areas of bionutrition research are most likely to yield tomorrow's breakthroughs?

Malnutrition is a global challenge to physiologists and nutritional biochemists as well as to public health and government agencies. Protein-energy malnutrition (PEM) — a nutritional deficiency illness that results in anaemia, poor growth, weakness and oedema — is its worst-case scenario. It occurs when the supply of protein, energy and micronutrients in the diet is simply insufficient to meet the body's needs for proper growth and development, or even for survival. If we can eliminate PEM, we can do the same thing for malnutrition in general.

PEM is a high-priority public health problem. A billion human beings are undernourished or malnourished in the world today^{1,2}. In North and South America alone, the 1995 estimate of infants with low birth weight was over a million; of girls and boys under the age of five, six million were seriously underweight as a result of the interaction between undernutrition and infection¹.

Insanitary environments breed repeated episodes of mild infectious illness or persistent chronic infections that impair the utilization of nutrients by the body and intensify the adverse impact of diets that lack both quality and variety. As well as the input from nutrition science, improvement in the general nutritional status of malnourished populations will require concerted public health efforts to improve water supplies, sanitation and immunization programmes.

Overcoming barriers

The prevention, diagnosis and treatment of PEM are highly dependent on the development of new technologies in food and nutrition science, new ways of delivering food to those who need it, and on changes in social, political and ecological systems (Fig. 1). Other factors that hamper progress include the complexities of malnutrition-related diseases, and the lack of scientific

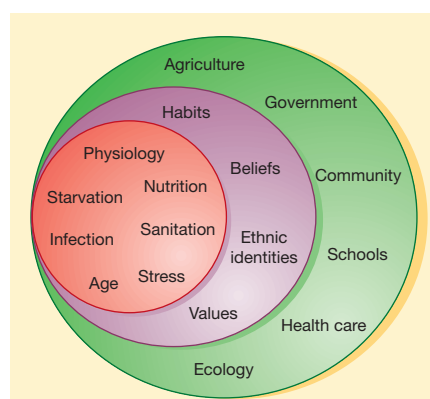


Figure 1 Framework for determinants of malnutrition. Prevention, diagnosis and treatment of PEM are highly dependent on new technologies in food and nutrition science, food-delivery systems, and changes in social, political and ecological systems.

insight into the damaging physiological changes that they cause.

These problems have impeded efforts to develop treatments for malnutrition-related diseases that are optimally tailored to particular populations and environments, for example, pregnant women who live in the poor, densely populated urban areas of Central America or children in remote, famine-stricken regions of Africa. Food and nutrition scientists, together with colleagues in related fields, can expedite success by revisiting assumptions about the relationship between basic and applied research, and by redesigning both treatments for malnutrition and the means of delivery to all those at risk. Malnutrition-related diseases are also diverse and often difficult to diagnose, with considerable potential for fatal errors in treatment. Given all these factors, scientists and clinicians will face many unforeseen challenges to successful intervention^{2,3}.

Pasteur's Quadrant

'Pasteur's Quadrant' is a modern framework for classifying the relationship between basic science and technological innovation. It is also a useful approach for exploring the necessary basic science of nutritional physiology that will lead to technologies for preventing and treating malnutrition. In his innovative book on science and technology policy⁴, the late Donald E. Stokes, professor of politics and public affairs in the Woodrow Wilson School of Public and International Affairs at Princeton University, moved away from a one-dimensional model of technological innovation stemming from basic research. He proposed a two-dimensional model in which innovations can be presented as paths within the quadrant between two axes: one representing developments in basic research, and the other developments in technology. In recasting the relationship between understanding and use in scientific research, Stokes cited the work of Louis Pasteur as a model of how fundamental yet use-inspired studies established the foundations of microbiology over a century ago. Since then, technology has grown ever more science-based. At the same time, science has become increasingly technology-based, with the choice of problems and the conduct of research often guided by societal needs.

Pasteur's Quadrant challenges the widely accepted validity of a dichotomy between basic and applied science. Taking inspiration from Pasteur, we need to look at the pathology and physiology of PEM to understand where the basic science of nutrition might help us most. For our purposes, Pasteur's Quadrant is best applied where the science of bionutrition — a term that refers to special opportunities afforded by biochemistry and food technology techniques that enrich nutrition science — intersects with the imperative to effectively address the many coexisting illnesses that result from malnutrition. New foods to prevent and treat these

Box 1 The physiology of starvation

Adequate clean water, energy, protein, vitamins (food-based organic substances that are not synthesized by the human body), minerals and certain other chemicals obtainable only from plants (for example, hormone-like substances such as carotenoids, flavonoids, lignans, sulphides and thiols) are necessary to sustain a balanced metabolism (metabolic homeostasis) and preservation of the body cell mass. Macronutrients are those nutrients required in relatively large amounts in the diet (amino acids, starches, sugars and fatty acids) and provide both the chemical building blocks from which cells are made and the metabolic energy that sustains life.

The healthy, well-fed person is protected from PEM by stores of glucose in the form of glycogen, labile protein reserves of amino acids, and a reserve of calories stored as fat. These can be mobilized during short periods without food and are built up again after a meal has been eaten. During such short periods of starvation, energy is obtained principally from free glucose and free fatty acids released from these stores¹³. Free glucose is the preferred fuel for such essential tissues as the blood-forming cells, the medulla of the kidney and the brain.

After three days without food, however, free glucose as an energy fuel becomes depleted. Glycogen stores in muscle and liver, which are used as primary energy sources, also become severely depleted. To continue to maintain blood glucose concentration, the liver starts to synthesize glucose from amino acids and other non-carbohydrates (gluconeogenesis).

During starvation, when glucose is unavailable, the primary fuels for brain, heart and muscle metabolism are ketone bodies, acetoacetate and β -hydroxybutyrate, all of which are derived from fatty acids. The breakdown of fat (lipolysis) in muscle and adipose tissue provides free fatty acids, the oxidation of which provides essential fuel and minimizes the use of visceral proteins for energy¹³. In kwashiorkor malnutrition, a deficiency of carnitine — an essential coenzyme in fatty acid metabolism that helps move fatty acid to the mitochondria from the cytoplasm — can impair the oxidation of fatty acids, leading to a state of advanced PEM.

Once a state of PEM is established, survival is threatened by reduced endogenous protein breakdown (proteolysis) coupled with depleted protein reserves and imbalances in the free amino acids available (Fig. 2). Many metabolic disorders that compromise cell function are the result of imbalances and deficiencies in amino acids. Cysteine, for example, is a component of many proteins and is also crucial for the synthesis of the tripeptide glutathione in red blood cells, which acts as a buffer for haemoglobin¹⁴.

illnesses will come from technological innovations rooted in basic research into the physiology of PEM⁵.

Protein–energy malnutrition

Although the first descriptions of PEM appeared in the 1800s, the condition was not recognized as a disease until the early twentieth century. The World Health Organization (WHO) and the Food and Agricultural Organization were founded in 1948. A year later they established clinical criteria for malnutrition and opportunities and priorities for future research.

In developing nations, PEM is most commonly found in overpopulated areas with scarce food supply or in environments blighted by famine, contaminated water and insanitary conditions. About 50% of children living in these conditions are malnourished or underweight, with symptoms of wasting and stunted growth. PEM coexists with hunger in areas of poverty and in groups with low socioeconomic status, where limited education means there is little understanding of diet and health. Severe malnutrition is also appearing in new adult populations, in particular in hospitalized patients with chronic diseases and in the institutionalized elderly in nursing homes, who cannot or will not eat. The mortality rate from severe malnutrition in these groups is 10 to 20 times higher than in healthy adults¹².

The physiology of PEM

PEM results when the body is stressed from starvation and is not receiving the protein and/or energy fuels and micronutrients needed to sustain the metabolic processes required for health and survival. The pathophysiology is failure of starvation and the related impetus of hunger to achieve the consumption of adequate amounts of essential foods. Prolonged starvation leading to severe PEM results in a cascade of physiological changes that cannot be resolved solely by the intake of available domestic foods. Therapeutic intervention requires nutrient-rich foods, oral rehydration fluids and often antibiotics. Functional foods could be important in promoting and sustaining recovery, particularly during the initial feeding period.

Functional foods are those that contain significant levels of biologically active components that impart health benefits or desirable physiological effects beyond basic nutrition. Functional attributes of many traditional foods are being discovered, while new food products are being developed to enhance or incorporate beneficial components. Ready-to-eat cereals are among the top ten food sources for 18 of 27 nutrients in the US diet, primarily because of fortification. Other examples include carotenoids and flavonoids to neutralize free radicals; ω -3 fatty acids to improve mental and visual functions; prebiotics/probiotics to improve the quality of

intestinal microflora; and sulphides and thiols to maintain immune function.

A wide range of clinical manifestations characterizes malnutrition and undernutrition; the relative intensity of the starvation, its duration and the age of the victim determine these. In particular, PEM is associated with the presence of other nutritional deficiencies, such as deficiencies of certain vitamins. It is also associated with infections and metabolic stresses such as are found in densely populated areas at risk of famine⁶.

There are two types of severe PEM — marasmus and hypoalbuminaemia or kwashiorkor. Marasmus is starvation that leads slowly to death, exemplified by the 'skin and bones' images of helpless infants displayed by relief organizations during aid appeals. In famine-stricken regions, it is the most common form of severe PEM, and is characterized by thinness, fatigue, irritability and weight loss of more than 20% of initial body weight. Victims have a low body-mass index or BMI — a measure used by the WHO as one method to categorize body size. BMI is calculated as the mass in kilograms (kg) divided by the square of the height in metres (m^2); a value below 19 kg m^{-2} is underweight. Sixty days of total starvation in a human of normal weight (7–12% body fat) results in marasmus malnutrition, which translates into a BMI $<16\text{ kg m}^{-2}$ in adults and a weight-for-height ratio in children of $<70\%$ of the original value. Terminal marasmus malnutrition presents with BMI values $<13.5\text{ kg m}^{-2}$ together with severe muscle wasting, gut atrophy and loss of subcutaneous fat.

Semi-starvation in children and adults, with consumption of less than 7 kcal per kg per day and less than 0.3 g protein per kg per day, can prolong the development of marasmus by months or even years. In such cases, survival depends on the quantity of fat stores, the degree of depletion of body cell mass, and the level of function in critical organ systems.

The main physiological changes that occur in starvation as the result of lack of macronutrients — starches, sugars, amino acids and fatty acids — are described in Box 1. In addition, the deficiencies in micronutrients, such as vitamins, trace elements and other compounds required in small amounts (phytonutrients), also have profound effects (Box 2).

As well as the symptoms of marasmus, malnutrition in children is characterized by a failure to grow. Amino acids are required for the biosynthesis of the protein hormones required for growth, such as insulin-like growth factor (IGF-1) and, in particular, IGF-1 binding protein 1. Such hormones are key control points at which nutritional status and metabolic homeostasis, the balanced functioning of essential metabolic processes, intersect.

Kwashiorkor develops in conditions of infection, diarrhoea, trauma and critical

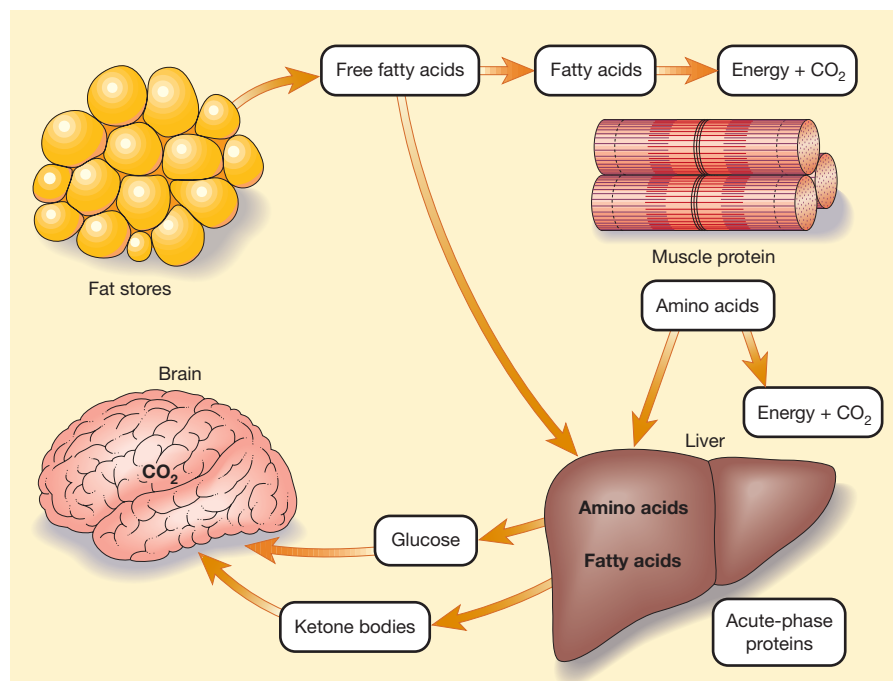


Figure 2 The metabolic response to malnutrition. Malnutrition triggers a functional redistribution of muscle protein (body cell mass) to provide the nitrogen necessary for synthesis of tissue protein, formation of red blood cells, wound healing and immune function (for example, the synthesis of acute-phase proteins). The interrelationship of fat, muscle and liver entails catabolism of muscle protein and release of gluconeogenic precursors for use by the liver. During starvation, ketone bodies are synthesized in the liver from fatty acids and used as fuel by the brain and other organs instead of glucose. The efficiency of physiological changes triggered by malnutrition is influenced by the extent of starvation and infection. It is important be aware of these changes when designing nutritional therapies.

illness, and occurs in some 2% of severely malnourished individuals. It is particularly prevalent in Africa, Central and South America and Southeast Asia, regions where the diet consists mainly of starchy vegetables contaminated with aflatoxin — a fungal toxin that commonly attacks crops in humid, temperate areas of the tropics.

In the developing world, children with oedematous kwashiorkor malnutrition often have their condition complicated by amino acid imbalances and various micronutrient deficiencies, including vitamins A and C and riboflavin⁷. They also suffer from tissue swelling that occurs when fluid leaks from the blood into the tissues, changes in the skin and hair, and low levels of secretory proteins such as serum albumin. Depending on the extent, onset and duration of the nutritional deficiency, they develop severe fluid retention and abdominal bloating, the swollen stomachs also familiar from images of famine. Shock and coma precede death.

Kwashiorkor malnutrition is physiologically more complex than marasmus⁸ because of the wide-ranging effects of selective deficiencies in amino acids, and its association with opportunistic diseases such as pertussis and tuberculosis. The energy- and protein-requiring immune response that the body mounts against infection compounds an

imbalance in the amino acids essential for protein synthesis and growth (Fig. 2).

Hormones and malnutrition

A shift from feasting to fasting alters the hormonal milieu in ways that signal a change in metabolic priorities from biosynthesis (anabolism) and growth to a post-absorptive state that maintains metabolic homeostasis between meals. During prolonged starvation, this hormonal response effects a redistribution of amino acids (from body muscle) and fatty acids (from adipose tissue) to meet the protein and energy needs of vital visceral-tissue functions of the brain, liver, intestine, bone marrow, heart and kidneys (Box 1). Hormonal modifications include reductions in insulin, IGF-1 and leptin, which act individually and together as anabolic hormones to muscle and adipose tissue. In contrast, there are increases in glucagon, glucocorticoids, growth hormone and catecholamines, which catabolize muscle (proteolysis) and adipose tissue (lipolysis) to meet the protein and energy requirements of visceral tissue.

Additional hormonal changes include increased production of reverse T3, the active tissue form of thyroid hormone, which results in decreased basal energy expenditure. The drop in basal metabolic rate is regulated primarily through a decline

in the activities of the thyroid and sympathetic nervous system.

The hormone leptin, produced by adipose tissue in response to starvation, is an important mediator between nutritional status and the neuroendocrine system. By its effects on the latter, leptin may prompt the maintenance of levels of cortisol and growth hormone necessary for effective lipolysis. During starvation, lipolysis is the only way to ensure an adequate supply of energy fuel for the brain (in the form of ketone bodies) and other organs (in the form of fatty acids).

Prolonged PEM leads to decreased synthesis of IGF-1 and low levels of insulin and leptin, or to their diminished effect owing to impaired receptor function in the presence of high levels of circulating growth hormone and cortisol. Their absence diverts the energy fuel substrate (that is, ketone bodies, acetoacetate and β -hydroxybutyrate, glutamine and other gluconeogenic amino acids) away from anabolism (biosynthesis and growth) and towards energy conservation and metabolic fuel adaptations that protect vital organs.

New 'functional foods'

A better understanding of the metabolic response to starvation and stress will be a key element in the evolution of strategies to develop new foods for the treatment of PEM.

The initial phase of feeding — for example, enteral delivery of nutrients and rehydration fluids — is particularly precarious in those with kwashiorkor or severe weight loss (>25% of usual or normal body weight). Feeding solutions need to meet the nutritional needs, stimulate appetite and cause minimal side effects. In cases of severe and prolonged starvation, feeding has to be slow (hourly), consistent and gradual to avoid clinical complications that include anorexia and re-feeding syndrome, a generalized fluid and electrolyte imbalance that may cause significant morbidity and mortality^{6,8}. If the patient regains his or her appetite within the first four days of the intervention, the feeding treatment can be considered a success.

Fatalities result when dietary therapy is limited only to macronutrient replacement. Indeed, inattention to micronutrients, salts and phytonutrients (for example: nutrients derived from phytochemicals; hormone-like substances found in plants, such as carotenoids, flavonoids, sulphides and thiols; ω -3 fatty acids; and probiotics) may explain the lack of correlation between severity of malnutrition and treatment outcome⁹.

Better treatment and prevention of PEM will require the development of specialized 'functional foods' — foods or food ingredients that provide a health benefit beyond that conferred by the nutrients the food contains¹⁰⁻¹². These functional foods will have to incorporate the micronutrients essential for increased protein synthesis, decreased protein catabolism and improved immune function

(see Box 2). Efforts to produce nutrient-fortified commodities and genetically modified (GM) foods containing vital micronutrients such as vitamin A and iron must also be based firmly in an understanding of the complexities of malnutrition. To achieve a complete re-alimentation diet, any therapeutic or functional food would have to supplement the intake of locally available foods.

As well as gross changes in physiology, many other potentially harmful changes occur in response to starvation. One of the key areas where intervention could be crucial is in the relatively little understood interactions between nutrition and the immune system, the body's defence against infection (see Box 2). Immune responses are essential for survival, but they are also vulnerable to protein, energy and micronutrient deficiencies. In malnourished individuals, immune responses generally tend to be suppressed, a state which enables opportunistic infections to flourish.

Food scientists should pay close attention to re-feeding diets rich in ω -6 fatty acids that increase production of prostaglandin E2 (PGE2). This compound is a potent inflammatory agent, and helps stimulate the development of immune responses. Properly designed ω -6/ ω -3 fatty acids can achieve the desired immune response and serve as potent antioxidants that can minimize the damage caused by free radicals. The presence of α -linolenic acid and other highly unsaturated fatty acids leads to inhibition of cytokine production, which in turn leads to suppression of the immune response.

The metabolic and endocrine adaptations to starvation are critical to survival during periods of extended nutritional deprivation. They cannot, however, minimize or slow the damage caused by multiple nutrient deficiencies linked to PEM; damage that underscores the need to prevent and treat this disease with carefully constructed diets or medical foods. A special focus is required on antioxidants, agents that minimize the negative impact of free radicals and lipid peroxidation and help maintain cell-mediated immunity. Their potential role in preventing organ failure, infectious disease and the morbidity, disability and mortality of PEM makes that abundantly clear.

Functional, nutrient-fortified and GM foods represent opportunities for the development of viable solutions to the problem of malnutrition. This applies not only to re-feeding for severe PEM, but also for cases of early-onset undernutrition — a condition with lifelong effects that could be reversed by changes in sociopolitical conditions (Fig. 1).

Unforeseeable technology

Bionutrition and functional foods

Those suffering from PEM need essential nutrients as well as specialized ones (conditional essential) that will reverse the effects of

Box 2 Micronutrients and malnutrition

PEM is also linked to shortfalls in micronutrients that are crucial in metabolism and in the proper functioning of the immune system. Imbalances and deficiencies in a variety of micronutrients, as well as amino acids and fatty acids, favour opportunistic infections, including tuberculosis, measles, malaria, hepatitis B (which leads to liver cancer), and secondary infections in HIV-compromised individuals.

Marginal micronutrient status also interferes with the efficiency of macronutrient utilization, further compromising the immune system. A diet restricted to very few foods can be lacking in some essential and conditionally essential amino acids. In general, a mixture of plant and animal proteins is required to optimize the use of macronutrients; the same applies to acute-phase proteins involved in host defence against infectious agents.

A diet deficient in antioxidants is associated with a rise in free radicals that cause tissue damage. The oedema of kwashiorkor, for instance, is linked to insufficient copper, zinc, selenium and vitamins A, E and C. These shortfalls, in turn, lead to increased production of superoxide dismutase, which results in lipid peroxidation and a rise in free iron, effects thought to be major causes of malnutrition-related oedema.

starvation, restore immune-defence mechanisms and recover body cell mass and energy stores. These nutrients, which change the hormonal milieu to an anabolism that facilitates re-feeding^{6-9,13,14}, could be supplied by functional foods¹⁰⁻¹².

Functional food ingredients can deliver nutrient density without overstimulating the hormonal response to the initial therapeutic feeding. They can be particularly effective in supporting organ function, protein synthesis and antioxidant activity required for early recovery. Foods fortified with vitamins C, E and A can modulate metabolic stresses and enhance immune status. Diets enriched with branched-chain amino acids, which are often provided in intravenous amino-acid solutions and enteral formula diets, can stimulate liver function, modulate insulin secretion and enhance cellular immune function. Diets fortified with fish oil, arginine and ribonucleotides can create nutritional immunomodulation that can minimize the metabolic stresses associated with malnutrition and infection. But the benefits that such immunostimulatory foods could confer would depend on a balanced total diet, adequate sanitation, and the availability of oral rehydration solutions, antibiotics and appropriate health care.

The health benefits of functional foods have never been as well developed as they are today. Bi-directional research, as embodied in the notion of Pasteur's Quadrant, will focus on the further development and refinement of functional foods and the investigation of their mechanisms of action. These efforts, in turn, will fuel growth in the field of bionutrition.

As nutrition scientists move into this new discipline, they will need to build on the wealth of information that already exists in plant biology. Many valuable natural phytonutrients are so-called 'secondary chemicals' produced by plants for purposes other than their general metabolism. The evolutionary and physiological bases of secondary chemical production in plants

will have to be considered to plan meaningful experiments that test the functionality of chemical compounds as foods for special dietary purposes.

Functional foods that possess gastrointestinal functions related to nutrient absorption, redox and antioxidant systems will be promising subjects for investigation, as will macronutrients that function in the re-feeding period of malnutrition treatment. But only a rigorous scientific approach that produces highly significant results will guarantee the success of bionutrition itself, and the development of foods for the prevention and treatment of PEM.

Nutrient cocktails

The metabolic response to stress is characterized in part by an increase in muscle catabolism, a reduction in total body protein synthesis, and increased oxidation of branched-chain amino acids (for example, novel essential amino acids, released from muscle, that stimulate liver protein synthesis). The result is an obligatory negative nitrogen balance (Fig. 2). Clinical investigation has led to the hypothesis that particular nutrients or mixes of nutrients may have a favourable impact on organ function independent of their general nutritional effects. Such 'nutrient cocktails' can reduce nitrogen loss and support protein synthesis better than standard amino-acid solutions¹¹.

In such cases, several normally dispensable amino acids become 'semi-essential' amino acids¹². Clinical studies show that supplemental dietary arginine, glutamate and tyrosine improve nitrogen retention and enhance immune response and brain neurotransmitter function. Designers of functional foods are thus currently investigating combinations of immunostimulatory nutrients such as branched-chain amino acids, arginine, glutamate, ribonucleotides and fish oil.

Structured lipids

The importance of polyunsaturated fatty acids (PUFA) in ameliorating inflammatory

disorders has long been recognized. ω -3 and ω -6 PUFA are precursors of eicosanoids and leukotrienes, both of which are involved in modulating inflammation. ω -3 fatty acids are present in certain plant oils and most fish oils. γ -Linolenic acid, a novel ω -6 fatty acid that also suppresses inflammation, is present in borage and evening primrose seeds. Although nature may not have provided the desired combination of fatty acids in a single triglyceride molecule for particular use as a functional food, this can be remedied by rearranging the properties of natural fatty acids to create new, structured lipids.

Structured lipids are created by enzymatically or chemically combining short-, medium- and long-chain fatty acids to form a new triglyceride molecule. These structurally mixed acids take advantage of two physiological pathways of absorption — long chains are absorbed through the lymph system, short and medium chains through the circulatory system. Specially synthesized fatty acids not found in nature, such as highly unsaturated ω -3 fatty acids, might be inserted into 2-glycerol, creating a molecule of unique digestibility and bioavailability.

New technology, such as innovations in computing power that enable instant processing and analysis of computationally intensive scientific and nutritional data, will enable the creation of new mixtures of structured lipids that can modulate the potentially damaging immune response that results from high levels of PGE₂.

Ligand-binding proteins

Other nutritional innovations include the use of ligand-binding proteins to stabilize volatile or potentially harmful agents (such as drugs, vitamins, flavourings or pheromones), shielding them from the environment or preventing their oxidation and degradation. Binding these agents to carriers with high affinity and selectivity for targeted sites can be useful in the fight against PEM, especially when subsequent controlled release can be engineered into these carriers. A protein envelope that acts as a diffusion barrier for the ligand can provide excellent slow-release features, which would be important in foods designed for therapeutic feeding.

Although natural proteins often possess one or more of the phytochemicals, biotechnology and bioinformatics provide the tools that bioengineers need to adapt these proteins to practical use. In fact, researchers have already engineered glucose-triggered release of glycosylated insulin from specially designed lectin carriers, and are using tools provided by biospecific antibodies to target nutrient, antioxidative and degradative compounds for special dietary purposes.

Probiotics and prebiotics

A probiotic is a live microbial dietary supplement that has a beneficial impact on the host

through its effects in the intestinal tract. Probiotics are available as freeze-dried cultures and are widely used to prepare fermented dairy products such as yoghurt. In the future they may also be found in fermented vegetables and meats. Several health-related effects associated with the intake of probiotics, including alleviation of lactose intolerance and immune enhancement, have been reported in human studies. Probiotics may be particularly valuable in reducing the risk of rotavirus-induced diarrhoea, which is a major source of malnutrition in infants in developing countries.

Prebiotics are non-digestible food ingredients that benefit the host by selectively stimulating the growth or activity of selected bacterial species that live in the intestinal tract. Inulin-type fructans, for example, are proving sufficiently promising to warrant their use as functional food ingredients.

The use of symbiotic or combined probiotics and prebiotics represents a promising direction for future research. Such combinations may be effective in improving the transit of viable, enriched microbial agents through the upper part of the gastrointestinal tract, thereby enhancing their delivery into the large bowel. This effect hastens the restoration of the normal gut flora at this site, and alleviates diarrhoea and dysentery.

Nutrient fortification

A classic approach to discovering the mechanisms by which diet prevents or treats malnutrition starts with the identification of individual nutrient components responsible for diet-related disease. This type of component research into single-nutrient, single-condition intervention will provide the rationale for developing nutrient-fortified and single-gene-modified foods.

Food technology

Processing methods can convert basic ingredients into practical foods and also change their physical and nutritional properties in useful ways. Extrusion cooking, for example, can process mixtures of plant foods into soup bases, flour mixes and other dry foods. The resulting products have functional properties that change their bulk density, water absorption and solubility indices. The use of extrusion cooking to add soybeans to sweet potatoes can increase the protein, fat, ash and trypsin inhibitor levels of the raw material mixtures.

New food technologies can also be used to speed and amplify the effect of drugs. Phytocompounds can be designed into bioavailability enhancers that modify drug metabolism in ways that increase the odds that the first dose will produce a strong therapeutic effect in any given individual. Phytocompounds also have the potential to alter the metabolism of pharmaceuticals and

prevent their elimination before they enter the blood stream. Reduction in the activity of drug-metabolizing enzymes is by far the most important of these effects.

Nutrition's pivotal role

Nutrition can be expected to play a pivotal role in optimizing health and productivity, and in reducing the factors that cause PEM in at-risk populations. The next generation of scientists will use complex systems and models to guide the development and evaluation of intervention programmes. The eradication of PEM and other forms of malnutrition will depend as much on their efforts as it will on the commitment of politicians, governments and transnational public health organizations to end the suffering caused by preventable and tragic nutrition-related disease.

Research outcomes can support recommended ways of preventing and treating PEM. The actual production, distribution and delivery of functional foods to those who need them, however, will require more than scientific effort alone.

Organizations such as the US National Institutes of Health's new Office of Dietary Supplements, the Agrotechnological Research Institute in The Netherlands, the Department of Food Science and Technology in Nigeria, and various academic institutions and industrial companies across the globe are currently providing leadership in the field of bionutrition^{10,15}. It will take these efforts and more to develop, test and realize the promise of tomorrow's breakthroughs.

George L. Blackburn is at the Beth Israel Deaconess Medical Center, 1 Autumn Street Kennedy 152, Boston, Massachusetts 02215, USA (e-mail: gblackbu@caregroup.harvard.edu).

1. World Health Organization. *The World Health Report 1999: Making a Difference* (WHO, Geneva, 1999).
2. UNU International Nutrition Foundation. *Ending Malnutrition by 2020: An Agenda for Change in the Millennium Final report to the ACC/SCN by the Commission on the Nutrition Challenges of the 21st Century* (eds James, P. J. et al.) <www.iotf.org/php/> (UNU International Nutrition Foundation, Boston, MA, 2000).
3. Scrimshaw, N. S. & SanGiovanni, J. P. *Am. J. Clin. Nutr.* **66**(Suppl.), 464S–477S (1997).
4. Stokes, D. E. *Pasteur's Quadrant — Basic Science and Technological Innovation* (The Brookings Institute, Washington DC, 1997).
5. Blackburn, G. L. *Am. J. Clin. Nutr.* **66**, 1067–1071 (1997).
6. World Health Organization. *Management of Severe Malnutrition: A Manual for Physicians and Other Senior Health Workers* (WHO, Geneva, 1999).
7. Waterlow, J. C. *Trans. R. Soc. Trop. Med. Hyg.* **78**, 436–441 (1984).
8. Collins, S., Myatt, M. & Golden, B. *Am. J. Clin. Nutr.* **68**, 93–99 (1998).
9. Golden, M. H. *Br. Med. Bull.* **54**, 433–444 (1998).
10. Harper, A. E. *Am. J. Clin. Nutr.* **71**(Suppl.), 1647S–1743S (2000).
11. Swails, W. S., Babineau, T. J. & Blackburn, G. L. in *Nutrient Substrates in Current Surgical Nutrition* (eds Latifi, R. & Dudrick, S. J.) 85–111 (Landes, Austin, Canada, 1996).
12. Mazza, G. *Functional Foods: Biochemical and Processing Aspects* (Technomic Publishing, Lancaster, PA, 1998).
13. Cahill, G. F. *Clin. Endocrinol. Metabol.* **5**, 397–415 (1976).
14. Reid, M. et al. *Am. J. Physiol. Endocrinol. Metabol.* **278**, E405–E412 (2000).
15. International Food Policy Research Institute (IFPRI) A 2020 Vision for Food, Agriculture, and the Environment: The Vision, Challenge, and Recommended Action (IFPRI, Washington DC, 1995).

Actions from thoughts

Miguel A. L. Nicolelis

Real-time direct interfaces between the brain and electronic and mechanical devices could one day be used to restore sensory and motor functions lost through injury or disease. Hybrid brain-machine interfaces also have the potential to enhance our perceptual, motor and cognitive capabilities by revolutionizing the way we use computers and interact with remote environments.

After a clever throw in by Tostão, a simple flick of Rivelino's magic left foot was enough to send the ball soaring into the thin air of the Azteca stadium in Mexico City. As the immaculate white object flew towards the middle of the penalty box on that hot afternoon, the colourful crowd that packed the stands slowly rose in anticipation. They roared, already celebrating, because they had seen that scene a thousand times before: the same graceful black man, dressed in blue shorts and a yellow jersey with the green 10 sewn in the back, defying logic, making fun of physics. The early celebration was warranted. As expected, Pelé floated above all Italian defenders to encounter the ball in mid air, and, with a gentle kiss of a forehead, changed its trajectory towards the net. Brazil had scored the first of its four goals in the final game of the 1970 World Cup and a whole country was about to start dancing in the streets.

The vast range of human abilities and behaviour illustrated here, as well as the gift of remembering the multitude of sensations associated with an instant of joy many decades ago, offer us a glimpse of the awesome repertoire of tasks that the human brain can accomplish. Through mechanisms that still elude our comprehension, the electrical activity of millions of brain cells (neurons) can be translated into precise sequences of skilled movements. Coordinated neuronal activity also provides us with exquisite perceptual and sensorimotor capabilities, illustrated in this example by Pelé's ability to track the ball's trajectory and plan the timing of his jump to hit it head on. But this is not all. Highly distributed patterns of neuronal firing underlie our ability to generate expectations about the outcome of a future event, learn the complex laws of nature and create art. One could argue, therefore, that hidden within the intricate principles that govern the way brain circuits operate lies the key to understanding the very essence of what it is to be human.

Witnessing the relentless growth of the disciplines that define modern neuroscience, one cannot help wondering what kind of insights, clinical applications and technologies may emerge from brain research in the future, and, more important, how they will impact on our lives. Although many of the imagined possibilities may not be feasible at this time, recent work indicates that some current ideas will come to fruition in the not-so-distant future. Here, I focus on one of these — the development of direct interfaces between machines and the human brain.

Brain-machine interfaces

I propose that the introduction of new methods for measuring large-scale brain activity, new techniques for microstimulating neuronal tissue, and emerging developments in microchip design, computer science and robotics have the potential to coalesce into a new technology devoted to creating interfaces between the human brain and artificial devices. One day, it is conceivable that such technology could allow patients to use brain activity to control electronic, mechanical or even virtual devices, leading to new therapeutic alternatives for restoring lost sensory, motor and even cognitive functions. Although many fundamental neurobiological questions and technical difficulties need to be solved, we can be optimistic about the feasibility of implementing this concept in the next few decades. Indeed, one brain-machine interface — the auditory prosthesis known as the cochlear implant — was introduced years ago and has improved the quality of life of many deafpatients^{1,2} (see Box 1).

Neuroscientists have long relished the possibility of using brain signals to control artificial devices³. As a consequence, there are already many terms in the literature⁴ to describe devices that could accomplish this goal (for example, brain-actuated technology, neuroprostheses or neurorobots). Here I will refer to these devices collectively as 'hybrid brain-machine interfaces (HBMI)s'. The word 'hybrid' reflects the fact that these applications rely on continuous interactions between living brain tissue and artificial electronic or mechanical devices.

My definition of HBMI's incorporates two main types of application. Type 1 devices use

artificially generated electrical signals to stimulate brain tissue in order to transmit some particular type of sensory information or to mimic a particular neurological function. The classic example of this application is an auditory prosthesis. Future applications aimed at restoring other sensory functions, such as vision, by microstimulation of specific brain areas would also belong to this group. In addition, type 1 HBMI's include methods for direct stimulation of the brain to alleviate pain, to control motor disorders such as Parkinson's disease⁵, and to reduce epileptic activity by stimulation of cranial nerves⁶. These last three applications rely on the observation that direct microstimulation of brain tissue can disrupt pathological patterns of brain activity that underlie some neurological disorders. Type 2 HBMI's rely on the real-time sampling and processing of large-scale brain activity to control artificial devices. An example of this application would be the use of neural signals derived from the motor cortex to control the

Box 1 Cochlear implants: the first HBMI

Auditory prostheses work by converting features of acoustic signals, such as speech, into patterns of electrical stimuli that are then delivered through an array of chronically implanted electrodes to auditory nerve fibres lying on the basilar membrane of the cochlea. As the basilar membrane contains a representation of sound frequencies, known as a tonotopic map, auditory prostheses deliver high-frequency information to the basal region of the cochlea, and low-frequency signals to the apical region, to mimic normal auditory processing. More than 30,000 deaf patients, ranging in age from 12 months to 80 years, have had such devices successfully implanted². Although results vary from case to case, even slight improvements in auditory performance have helped people to communicate better and to become more aware of their surrounding environment.

movements of a prosthetic robotic arm in real time. Obviously, clinical applications that require reciprocal interaction between the brain and artificial devices will combine both type 1 and 2 HBMI.

The design and implementation of HBMI will involve the combined efforts of many areas of research, such as neuroscience, computer science, biomedical engineering, very large scale integration (VLSI) design and robotics. I have selected a few current developments in these fields to illustrate below some of the conceptual advances and technologies that will be required to design and implement useful HBMI. I will then describe two potential clinical applications of such technology that should emerge in the near future: a system to monitor and treat epileptic seizures and a device to control a robotic prosthetic arm.

Building a HBMI

The first of the many challenges associated with the development of any HBMI is the need to understand better the principles by which neural ensembles encode sensory, motor and cognitive information. This is rapidly becoming one of the main goals of modern neuroscience, but our present knowledge is elementary at best. In the case of motor control, for instance, the areas of the primate brain involved are well known, and considerable information is available on the physiological properties of individual neurons located in each of them. But we know little about how the brain makes use of information from these neurons to generate movements. To design a type 2 HBMI that uses brain-derived signals to control a prosthetic robotic arm, we will need to learn how to sample and decode the motor signals generated by neurons and how to feed them into an artificial device to mimic the intended movement.

Recording brain activity

It is clear that neurobiological principles will be central in devising a strategy to overcome these hurdles. For example, classic experiments in primates have demonstrated that fundamental parameters of motor control emerge by the collective activation of large distributed populations of neurons in the primary motor cortex (M1). Single M1 neurons are broadly tuned to the direction of force required to generate a reaching arm movement⁷. In other words, even though these neurons fire maximally before the execution of a movement in one direction, they also fire significantly before the onset of arm movements in a broad range of other directions. Therefore, to compute a precise direction of arm movement, the brain may have to perform the equivalent of a neuronal 'vote' or, in mathematical terms, a vector summation of the activity of these broadly tuned neurons⁷. This implies that to obtain the motor signals required to control an artificial

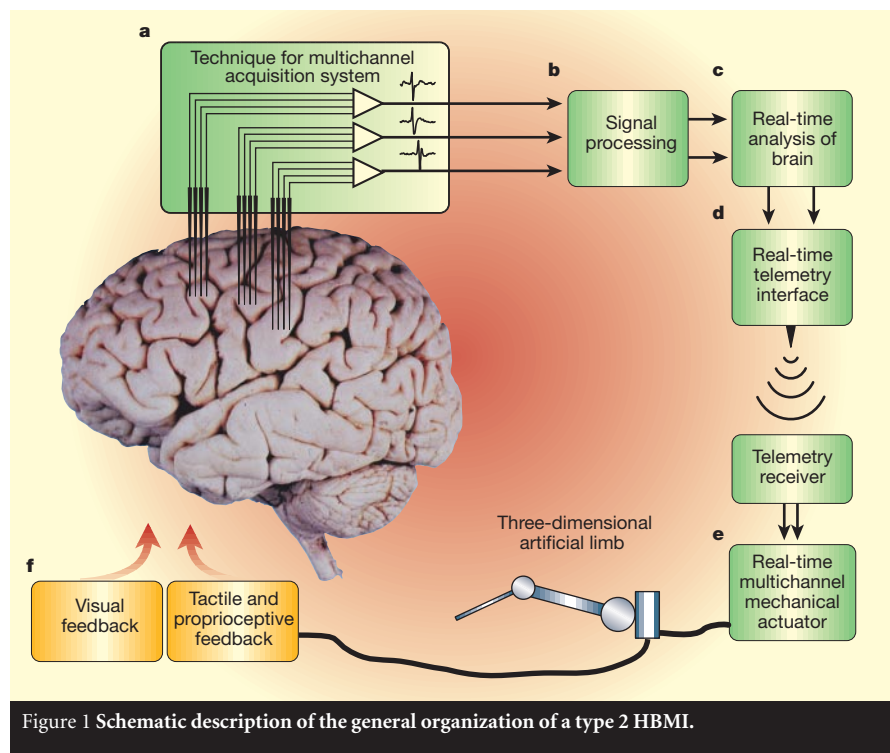


Figure 1 Schematic description of the general organization of a type 2 HBMI.

device we will need to sample the activity of many neurons simultaneously and design algorithms capable of extracting motor control signals from these ensembles. Moreover, it will be crucial to investigate how these neural ensembles interact under more complex and 'real-world' experimental conditions⁸ to generate different motor behaviours. These data will be vital to answering basic questions in regard to the development of type 2 HBMI. For example, what is the minimum neuronal sample required to generate reliable brain-derived control signals? Should these samples be obtained from one or multiple brain areas? Does the same population of neurons code for single or multiple control parameters? Finally, how might neural encoding mechanisms change with time, experience and learning?

Figure 1 illustrates the general organization of a type 2 HBMI and depicts some of the technological challenges involved in designing such devices. The first design step involves the selection of a technique (Fig. 1a) that yields reliable, stable and long-term recordings of brain activity that can be used as control signals to drive an artificial device. From recent studies in animals^{9,10}, clinical applications of HBMI will probably require sampling of large numbers of neurons (in the order of hundreds or thousands) with a temporal resolution of 10–100 ms, depending on the application.

Although neuroscientists have long recognized the need to investigate the properties of large neural ensembles¹¹, it is very difficult to obtain reliable, long-term measurements of neural ensemble activity with high spatial and temporal resolution. Starting in the

1940s and 50s with multichannel recordings of scalp electroencephalographic (EEG) activity and of the general electrical activity evoked by movement or sensory stimulation, a variety of metabolic, optical and electrophysiological methods have been introduced for monitoring large-scale brain activity. Modern multichannel electrophysiological recordings are made from arrays of microelectrodes surgically implanted in the brain. They currently allow neurophysiologists to record simultaneously, with a resolution of milliseconds, the extracellular activity of up to 100 individual neurons, distributed across multiple brain structures, in animals carrying out some task or other¹². Although future improvements might allow long-term and non-invasive sampling of human neural activity with the same temporal resolution as intracranial recordings, the first generation of HBMI will probably rely on improved versions of electrophysiological methods, such as multichannel EEG or multielectrode intracranial recordings. Indeed, preliminary studies in paralysed patients have shown that EEG signals can be used to trigger the movement of computer cursors¹³ or offer a way for patients to communicate¹⁴.

Unfortunately, less invasive electrophysiological methods, such as scalp EEG recordings that reflect the common electrical activity of millions of neurons in widespread areas of the cortex, lack the resolution to provide the kind of time-varying motor signals needed to control a robotic arm in real time⁴. Multichannel intracranial recordings of brain activity, obtained by surgical implantation of arrays of microwires within one or more cortical motor areas, will

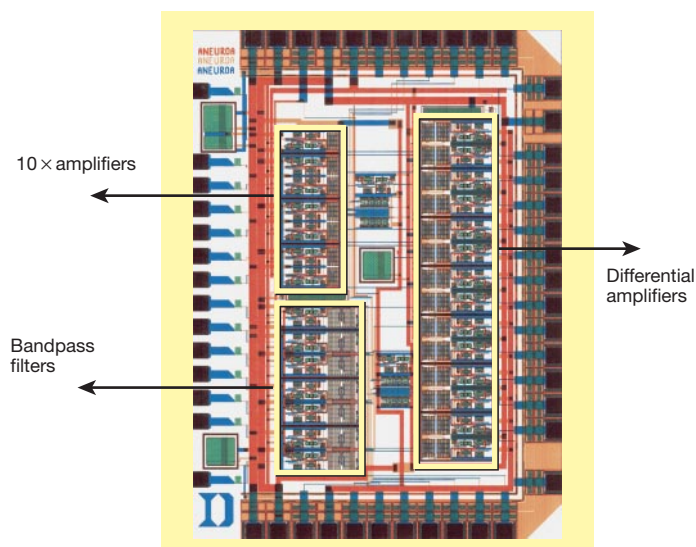


Figure 2 A prototype of an instrumentation neurochip for processing brain-derived signals. This chip, containing a portion of the analog signal processing for 16 neural channels, was designed by I. Obeid, H. Aurora, J. Morizio and P. Wolf in the Departments of Biomedical and Electrical & Computer Engineering at the Pratt School of Engineering, Duke University. The mixed-signal CMOS (complementary metal-oxide semiconductor) process used in the design supports digital signal processing modules which will be included in future generations of this device.

therefore be required, with mathematical analysis of the extracellular activity of smaller populations (100–1,000) of neurons providing the raw brain signals for use in most HBIMs¹⁰. Despite some degree of recording degradation over time, present technology allows simultaneous sampling of 50–100 neurons, distributed across multiple cortical areas of small primates, to remain viable for several years^{10,12}. Technological advances in multielectrode array design and neural signal instrumentation in the next decade alone are expected to increase the number of neurons that can be recorded simultaneously by at least one order of magnitude.

The precise placement of the electrode arrays for intracranial recording may not be as critical to the ability to control an artificial device as was first conjectured. As motor control signals emerge from the distributed activation of large populations of neurons, and as cortical and subcortical neurons are capable of considerable plastic reorganization during adulthood¹⁵, electrode arrays targeted to brain areas of interest may suffice in most cases. As subjects learn to interact with artificial devices through HBIMs, it is likely that sampled neurons that were not originally involved in the type of motor control to be mimicked may be recruited into generating the signals required to control artificial devices.

Generating the output

After selecting a method for acquiring the necessary brain signals, the next challenge is to design the instrumentation (Fig. 1b–d) required for recording and processing these

signals in real time. Currently, this requires specialized, sizeable and expensive electronic equipment, which can amplify and filter the original signal as well as perform analog-to-digital conversion to facilitate further processing and storage of data. To make HBIMs viable, new technologies for portable, wireless-based, multichannel neural signal instrumentation are needed.

The central issue of signal conditioning and instrumentation may be solved in the near future by the application of mixed-signal VLSI into the design of neurophysiological instrumentation chips. This technology allows analog and digital signals to coexist in the same microchip, and has the potential to provide the multichannel, programmable and low-noise package required for conditioning brain-derived signals for clinical implementation of HBIMs. Moreover, the resulting microchip would be small enough to be chronically implanted in patients and could be powered by replaceable batteries. Such microchips could rely on wireless communication protocols based on a radio frequency link to broadcast neural signals to other components of the HBIM (Fig. 1d–e).

Prototypes of dedicated ‘instrumentation neurochips’ (Fig. 2) are currently being developed, although many complex issues must be solved before they can become operational¹⁶. For instance, efficient solutions will have to be found to provide enough power for performing analog and digital processing, while still ensuring that signals can be transmitted by telemetry. Thus, battery technology, device packing and the bandwidth of the neural signals, among

other factors, will certainly be important in the design of HBIMs¹⁶.

Having selected a method for sampling and conditioning brain signals, the next step—and one of the most difficult challenges—is to define a strategy for extracting meaningful control information from neural ensemble activity in real time. Currently, neuroscientists rely on a variety of linear and nonlinear multivariate algorithms, such as discriminant analysis, multiple linear regression and artificial neural networks, to carry out real-time and off-line analysis of neural ensemble data. Preliminary results from animal studies that use these different methods are encouraging, but considerably more experience is needed to apply these techniques in clinical HBIMs. The challenge is to produce algorithms that can combine the activity of large numbers of neurons, which convey different amounts of information, and extract stable control signals, even when the firing patterns of these neurons change significantly across different timescales. Research on areas ranging from automatic sorting algorithms for unsupervised isolation of single neuron action potentials, to the design of real-time pattern recognition algorithms that can handle data from thousands of simultaneously recorded neurons will certainly be required. In the same context, clinical applications of HBIMs will require considerable computational resources.

In the not too distant future, new developments in the design of brain-inspired VLSI¹⁷, an exciting area of research aimed at modelling neuronal systems in silicon¹⁸, may provide the means for achieving the type of efficient real-time neural signal analysis required for HBIMs. This technology may allow pattern recognition algorithms, such as artificial neural networks or realistic models of neural circuits, to be implemented directly in silicon circuits. Among many other technical hurdles, significant work will be required to make these silicon circuits adaptive, perhaps by incorporating learning rules derived from the study of biological neural circuits. This will allow ‘training’ of algorithms as well as ensuring the robustness of the control system. From an implementation point of view, ‘analytical neurochips’ are ideal as they could be interfaced with the instrumentation neurochip and be chronically implanted in the subject.

The final component of the idealized HBIM (Fig. 1e–f) is a real-time control interface which uses processed brain signals to control an artificial device. The types of devices used are likely to vary considerably in each application, ranging from elaborate electrical pattern generators to control muscles, to complex robotic and computational devices designed to augment motor skills¹⁹.

HBIMs for epilepsy control

Estimates indicate that about 0.5–2.0% of the population has epilepsy²⁰. About 10–50% of these patients do not respond well to current

antiepileptic medications and may not be candidates for surgery. Throughout this century, neuroscientists have used multi-channel recording from scalp, brain surface and even chronically implanted intracranial electrodes to investigate the electrophysiological activity that characterizes different types of seizure in humans. By doing so, scientists have not only identified different types of epilepsy, but they have also learned that there are distinct patterns of neurophysiological activity associated with the initiation and establishment of a seizure attack.

Several exciting new developments in epilepsy research indicate that the development of an unsupervised HBMI for monitoring, detecting and treating seizure activity may be possible in the next decade (Fig. 3a). First, for certain types of seizure, there seems to be a particular spatiotemporal pattern of cortical activity that appears seconds or even minutes before the full epileptic attack starts²¹. Recently, a few laboratories have introduced automatic seizure-prediction algorithms that can be applied to intracranial and scalp recordings to forecast the occurrence of a seizure^{21,22}. These and future seizure-prediction algorithms might provide sufficient time (2–5 minutes) to warn the patient of an imminent attack, and to trigger automatic therapeutic intervention before convulsion or loss of consciousness.

But what kind of therapy could be triggered that would work in patients who are refractory to epilepsy medication? The answer may lie in another recent development in epilepsy research. Studies in both animals²³ and human subjects⁶ have revealed that electrical stimulation of peripheral cranial nerves, such as the vagus²³ and trigeminal²⁴ nerves, can substantially reduce cortical epileptic activity. Moreover, if this peripheral nerve stimulation is applied before the initiation of seizure or during its initial stages, significantly higher reduction of seizure activity can be achieved.

From this I believe that a device containing a combination of both type 1 and 2 HBMI could be designed to function somewhat like a modern heart pacemaker (Fig. 3a). This 'brain pacemaker' would rely on arrays of chronically implanted electrodes to search continuously for spatiotemporal patterns of cortical activity indicating an imminent epileptic attack. Instrumentation neurochips would be responsible for all the basic signal-processing operations. They would also provide signals to one or more seizure-prediction algorithms, implemented into analytical neurochips, which would carry out real-time analysis of cortical activity. Once pre-seizure activity patterns were detected, the analytical neurochip could trigger electrical stimulation of one or multiple cranial nerves. In patients who respond to pharmacological therapy, the same stimulator could be used to activate a mini-pump to deliver one or more anti-epileptic

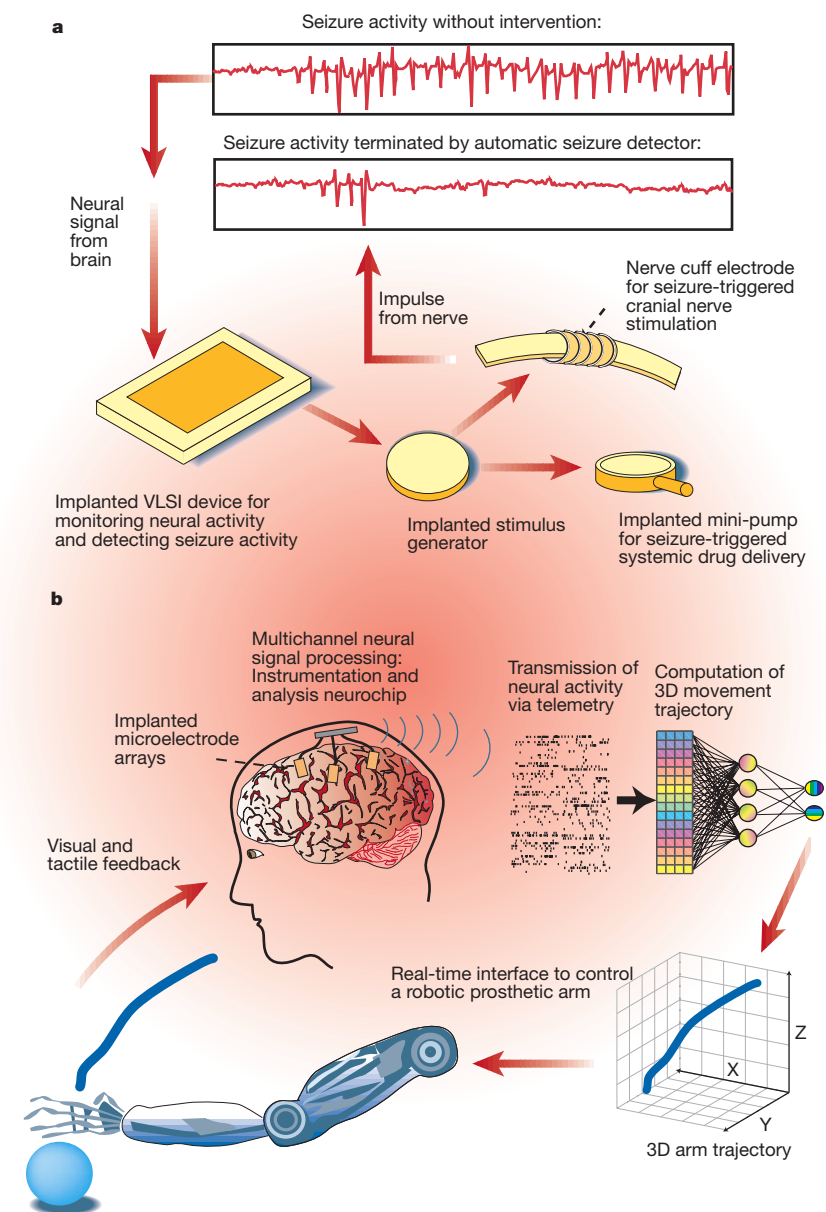


Figure 3 Schematic description of two potential applications of type 2 HBMI. **a**, Design of a 'brain pacemaker' that monitors neural activity using a VLSI chip designed to detect seizure activity. When seizure activity is detected, the VLSI chip sends a signal to an implanted stimulus generator that drives either a nerve cuff electrode or a mini-pump for drug delivery, either of which can stop the seizure activity. **b**, HBMI for controlling a robotic prosthetic arm using brain-derived signals. Multiple, chronically implanted, intracranial microelectrode arrays would be used to sample the activity of large populations of single cortical neurons simultaneously. The combined activity of these neural ensembles would then be transformed by a mathematical algorithm into continuous three-dimensional arm-trajectory signals that would be used to control the movements of a robotic prosthetic arm. A closed control loop would be established by providing the subject with both visual and tactile feedback signals generated by movement of the robotic arm.

drugs directly into the blood stream. Recently, a simplified implementation of this concept has been used successfully in rats²⁴, giving hope that a brain pacemaker for seizure monitoring and control in humans may not be far ahead.

HBMI to restore motor function

Another clinical application of HBMI that could emerge in the near future aims at

restoring different aspects of motor function in patients with severe body paralysis, caused primarily by strokes, spinal cord lesions or peripheral degenerative disorders (Fig. 3b). Advances in this rapidly growing field of research indicate that neural signals from healthy regions of the brain could be used to control the movements of artificial prosthetic devices, such as a robotic arm. Preliminary

findings also demonstrate that paralysed patients can learn to use brain signals obtained from their motor cortex to interact with computers²⁵.

Extensive electrophysiological work in primates and imaging studies in humans have shown that multiple interconnected cortical areas in the frontal and parietal lobes are involved in the selection of motor commands that control the production of voluntary arm movements²⁶. Although each of these areas has different degrees of functional specialization, in theory, each of them could be selected as the source of brain signals for controlling the movements of an artificial device. Within each of these cortical areas, different motor parameters, such as a force and direction of movement, are coded by the distributed activity of populations of neurons, each of which is typically broadly tuned to one (or more) of these parameters. This indicates that implementations of HBMs for robotic arm control need to rely on intracranial recordings from large populations of single neurons to derive motor control signals.

At a first glance, a random sample of 100–1,000 cortical motor neurons, which represents a reasonable expectation for the yield of multielectrode intracranial recordings in the near future, may look too small to unveil any useful information. But recent neurophysiological experiments dispute this view. For instance, currently one can obtain precise off-line reconstructions of complex three-dimensional arm trajectories by using simple multiple regression techniques to transform the activity of 300–400 serially recorded cortical motor neurons into a neural population vector²⁷. Moreover, experiments in rats⁹ and primates¹⁰ have shown that simple, real-time algorithms, applied to samples of 50–100 simultaneously recorded cortical neurons, can be used to control robotic devices in real time and mimic the type of three-dimensional arm reaching movements produced by primates.

Another important issue is that, to achieve seamless interactions with prosthetic devices, patients will have to receive sensory feedback information (for example visual or tactile signals) from the prosthetic limbs. These feedback signals will establish a closed control loop between the brain and artificial devices and will probably help patients learn how to operate HBMs. Studies in rats have revealed that, if subjects receive visual feedback information as they learn to use brain activity to interact with a robotic arm, and are rewarded for the successful completion of these movements, they progressively cease to produce overt limb movements⁹. In other words, even though the rats continued to exhibit the patterns of cortical activity required to control the movements of the robotic arm, this motor activity did not result in any significant limb movement. This indicates that motor control signals can be generated by cortical neurons

without any muscle activity, and hence that paralysed patients might be capable of learning to operate a robotic arm even though they cannot move their own limbs.

These observations also raise the intriguing hypothesis that, by establishing a closed control loop with an artificial device (Fig. 3b), the brain could incorporate electronic, mechanical or even virtual objects into its somatic and motor representations, and operate upon them as if they were simple extensions of our own bodies. The fact that the adult cortex is capable of significant functional reorganization (or plasticity) after peripheral and central injuries¹⁵, changes in sensory experience²⁸ and learning of new motor skills²⁹ supports this possibility. Indeed, the notion that adult plasticity can dynamically alter the perception of the limits of our own body is corroborated by studies on patients who have undergone limb amputations. Immediately after the amputation, most of these patients experience the sensation that their amputated limb is still present and moving. These 'phantom limb' sensations are paralleled by a significant plasticity of body maps in the somatosensory cortex³⁰, the part of the brain that receives and interprets sensory signals from areas such as the skin surface. Instead of remaining silent, the areas in these brain maps that used to represent the amputated limb progressively start to respond to stimulation of neighbouring body regions spared by the amputation. Thus, it is conceivable that tactile feedback signals, generated by the movements of a brain-controlled robotic arm and delivered to the patient's skin, could be used to incorporate the representation of such an artificial device into cortical and subcortical somatotopic maps.

Undoubtedly, years of research will be required and many fundamental technological breakthroughs needed before this comes close to reality. Nevertheless, it seems reasonable to predict that a definitive demonstration of such a phenomenon could trigger a revolution in the way future generations interact with computers, virtual objects and remote environments, by allowing never-before-experienced augmentation of perceptual, motor and cognitive capabilities. Such applications, however, will require the introduction of new, non-invasive methods for sampling brain activity.

A final thought

Although developing expectations of a distant future is a risky business and may raise unjustified hope that solutions are just around the corner, I cannot avoid ending this brief overview on an optimistic tone. Despite many significant conceptual and technological obstacles, the possibility of developing clinical applications of HBMs is real and worth pursuing, especially given the potential benefits that they may bring to people afflicted by neurological disorders. At the very least,

research on HBMs will yield powerful new tools to investigate hypotheses of how large populations of neurons process information and adapt according to changes in experience. Some may argue that one could achieve this goal just by building theoretical models and running computational simulations. Perhaps that is true. But as my good friend Idan Segev, a leading computational neuroscientist, always tells me, there is a subtle but fundamental difference between simulating reality and building it. Those of us who saw Pelé scoring that magic goal on that hot Mexican afternoon in 1970 and dreamed about doing the same thing would certainly agree.

Miguel A. L. Nicolelis is in the Departments of Neurobiology, Experimental Psychology, and Biomedical Engineering, Duke University, Durham, North Carolina 27710, USA

(e-mail: nicoleli@neuro.duke.edu)

- Merzenich, M. M., Schindler, D. N. & White, M. W. *Laryngoscope* **84**, 1887–1893 (1974).
- Pfingst, B. E. in *Neural Prostheses for Restoration of Sensory and Motor Function*. (eds Chapin, J. K. & Moxon, K. A.) 3–43 (CRC, Boca Raton, 2000).
- Schmidt, E. M. *Ann. Biomed. Eng.* **8**, 339–349 (1980).
- Chapin, J. K. & Moxon, K. A. (eds) *Neural Prostheses for Restoration of Sensory and Motor Function*. (CRC, Boca Raton, 2000).
- Benabid, A. L. *et al. Lancet* **337**, 403–406 (1991).
- Uthman, B. M., Wilder, B. J., Hammond, E. J. & Reid, S. A. *Epilepsia* **31**(Suppl. 2), S44–S50 (1990).
- Georgopoulos, A. P., Swartz, A. B. & Ketter, R. E. *Science* **233**, 1416–1419 (1986).
- Ghazanfar, A. A. & Hauser, M. D. *Trends Cog. Sci.* **3**, 377–384 (1999).
- Chapin, J. K., Moxon, K. A., Markowitz, R. S. & Nicolelis, M. A. L. *Nature Neurosci.* **2**, 664–670 (1999).
- Wessberg, J. *et al. Nature* **408**, 361–365 (2000).
- Hebb, D. O. *The Organization of Behaviour: A Neuropsychological Theory* (Wiley, New York, 1949).
- Nicolelis, M. A. L. *et al. Nature Neurosci.* **1**, 621–630 (1998).
- Wolpaw, J. R., McFarland, D. J., Neat, G. W. & Forneris, C. A. *Electroencephalogr. Clin. Neurophysiol.* **78**, 252–259 (1991).
- Schutz, S. *et al. Nature* **398**, 297–298 (1999).
- Wu, C. W. & Kaas, J. H. *J. Neurosci.* **19**, 7679–7697 (1999).
- Moxon, K. A., Morizio, J., Chapin, J. K., Nicolelis, M. A. L. & Wolf, P. D. in *Neural Prostheses for Restoration of Sensory and Motor Function*. (eds Chapin, J. K. & Moxon, K. A.) (CRC, Boca Raton, 2000).
- Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J. & Seung, H. S. *Nature* **405**, 947–951 (2000).
- Mead, C. *Analog VLSI and Neural Systems* (Addison-Wesley, Reading, MA, 1989).
- Srinivasan, M. A. in *In Virtual Reality: Scientific and Technical Challenges* (eds Durlach, N. I. & Mavrou, A. S.) 161–187 (National Academy Press, 1994).
- McNamara, J. O. *Nature* **399**(Suppl.), A15–A22 (1999).
- Martinerie, J. *et al. Nature Med.* **4**, 1173–1176 (1998).
- Webber, W. R. S., Lesser, R. P., Richardson, R. T. & Wilson, K. *Electroencephalogr. Clin. Neurophysiol.* **98**, 250–272 (1996).
- Zabara, J. *Epilepsia* **33**, 1005–1012 (1992).
- Fanselow, E. E., Reid, A. P. & Nicolelis, M. A. L. *J. Neurosci.* **20**, 8160–8168 (2000).
- Kennedy, P. R. & Bakay, R. A. *NeuroReport* **9**, 1707–1711 (1998).
- Wise, S. P., Boussaoud, D., Johnson, P. B. & Caminiti, R. *Annu. Rev. Neurosci.* **20**, 25–42 (1997).
- Schwartz, A. *Science* **265**, 540–542 (1994).
- Polley, D. B., Chen-Bee, C. H. & Frostig, R. D. *Neuron* **24**, 623–637 (1999).
- Laubach, M., Wessberg, J. & Nicolelis, M. A. L. *Nature* **405**, 567–571 (2000).
- Ramachandran, V. S. *Proc. Natl. Acad. Sci. USA* **90**, 10413–10420 (1993).

Acknowledgements. I thank J. Chapin, F. Ebner, E. Fanselow, A. Ghazanfar, C. Henriquez, J. Kaas, J. Kralik, D. Krupa, S. Ribeiro, V. de Sa, I. Segev, M. Shuler, S. Simon, J. Wessberg and P. Wolf for comments and suggestions, and E. Fanselow, D. Krupa, P. Beck and J. Wessberg for their help in creating the illustrations for this article. The author's research on HBM is funded by NIH, DARPA and ONR grants.

The relationship between matter and life

Rodney Brooks

The disciplines of artificial intelligence and artificial life build computational systems inspired by various aspects of life. Despite the fact that living systems are composed only of non-living atoms there seems to be limits in the current levels of understanding within these disciplines in what is necessary to bridge the gap between non-living and living matter.

Researchers in artificial intelligence (AI) and artificial life (Alife) are interested in understanding the properties of living organisms so that they can build artificial systems that exhibit these properties for useful purposes. AI researchers are interested mostly in perception, cognition and generation of action (Box 1), whereas Alife focuses on evolution, reproduction, morphogenesis and metabolism (Box 2). Neither of these disciplines is a conventional science; rather, they are a mixture of science and engineering. Despite, or perhaps because of, this hybrid structure, both disciplines have been very successful and our world is full of their products.

Every time we use a computer we use algorithms and techniques developed by AI researchers. These range from the natural language processing and indexing techniques in web search engines to the bayesian matching techniques used in help and document autoformatting systems in our word processors. When we play a video game our opponent is usually an AI system. At many airports, an AI program schedules our arrival gate, and when we apply for credit an AI neural network often vets our application. When we watch a film with digitally generated crowds, be they aliens or ants, we are watching groups of agents acting under Alife models of group behaviour. When we fly in the latest aeroplane, the design of the turbines may have been optimized by artificial evolution.

But despite all this, both fields have been labelled as failures for not having lived up to grandiose promises. At the heart of this disappointment lies the fact that neither AI nor Alife has produced artefacts that could be confused with a living organism for more than an instant. AI just does not seem as present or aware as even a simple animal and Alife cannot match the complexities of the simplest forms of life.

Moore power...

Part of the problem was the lack of computer power in the early years of AI and Alife.

Moore's law states that computational resources for a fixed price roughly double every 18 months. From about 1975 into the early 1990s all the gains of Moore's law went into the changeover from the centralized mainframe to the individual computer on your desk, accommodating a vastly increased number of users. The amount of computing power available to the individual scientist did not change that much, although the price came down by a factor of a thousand. But since the early 1990s, all of Moore's law has gone into increasing the performance of the workstation itself. And both AI and Alife have benefited from this shift.

Increased computer power has enabled search-based AI to push ahead with programs that achieve their ends through brute force — the Deep Blue program that beat the world chess champion is a good example. The essential ideas were in place in Greenblatt's 1965 program MacHack¹, but this could process only a few thousand possible chess moves per second. By 1997, when Deep Blue beat Kasparov, it was processing 200 million moves per second. More power has also enabled implementation of real-time perceptual systems, often based on neural models, that can simulate in serial computers the massive parallelism found in the brains of animals. Marr and Hildreth² required 10 minutes of computer time to find the edges in a single image in the late 1970s; we now have computer vision systems that track multiple moving objects in a scene at 30 frames per second. Others can visually track the boundaries of roads and cars a few times per second. Using such a system, the 'No hands across America'³ project, at Carnegie Mellon University, made an automated truck drive from the east to the west coast of the United States with no human control for 98% of the journey.

Much more complex systems⁴ can now be modelled as Alife, down to the level of molecules and enzyme-like components interacting to produce aspects of life, although the

complexity of the models is still far below that of any living system. New experiments in evolution simulate spatially isolated populations to investigate speciation. Over the past few years, new directions have emerged in AI⁵, in attempts to implement artificial creatures in simulated or physical environments.

Often called the behaviour-based approach, this new mode of thought involves the connection of perception to action with little in the way of intervening representational systems. Rather than relying on search, this approach relies on the correct short, fast connections being present between sensory and motor modules. Behaviour-based approaches began with insect models, but more recently they have been extended to humanoid robots⁶ — robots with human form that can interact with people in a social manner, eliciting natural and involuntary social responses from naive subjects.

Behaviour-based systems are now making their mark in consumer products such as the new generation of intelligent dolls, but perhaps their greatest success was the Sojourner Mars rover⁷. For the final phase of its mission in 1997 this behaviour-based robot was allowed to operate autonomously, and successfully navigated the surface of Mars.

The behaviour-based approach to AI has merged somewhat with the Alife endeavour, and a community of researchers has formed that is separate from the traditional AI community. The former is interested in understanding how living systems work and in building computational and physical models of them. The latter is interested in building systems with maximal performance, and is usually wary of biological inspiration as taking away from mathematically optimized engineering solutions.

Problems, problems, problems

Although they are much more lifelike than the pure engineering artefacts of traditional AI, in some sense the systems built under the behaviour-based and Alife approaches do

not seem as alive as we might hope. We build models to understand the biological systems better, but the models never work as well as biology. We have become very good at modelling fluids, materials, planetary dynamics, nuclear explosions and all manner of physical systems. Put some parameters into a program, let it crank, and out come accurate predictions of the physical character of the modelled system. But we are not good at modelling living systems, at small or large scales. Something is wrong.

Solutions and new developments?

What is going wrong? There are a number of possibilities: (1) we might just be getting a few parameters wrong; (2) we might be building models that are below some complexity threshold; (3) perhaps it is still a lack of computing power; and (4) we might be missing something fundamental and currently unimagined in our models of biology.

Incorrect parameters

Getting just a few parameters wrong would mean that we have essentially modelled everything correctly, but are just unlucky or ignorant in some minor way. With a bit more work on our part, things will start working better. It could be that our current neural-network models will work quantitatively better if we have five layers of artificial neurons, rather than today's standard of three. Or that artificial evolution works much better with populations of 100,000 or more, rather than the typical thousand or less. But this seems unlikely. One would expect that someone would have stumbled by now across a combination of parameters that worked qualitatively better than anything else around. That success would have led to theoretical analysis and we would have already seen rapid progress.

Models lack complexity

Building models that are below some complexity threshold also would mean that there is nothing in principle that we do not understand about intelligent or living systems. We have all the ideas and components lying around, we just have not yet put enough of them together in one place, or one model. When, and if, we do, then everything will start working a lot better. As for the first possibility, while this may be true, it does seem unlikely that is true across so many different aspects of biology.

A lack of computing power

We have recently seen an example of this. After being defeated by Deep Blue, Garry Kasparov said that he was surprised by its "ability to play as though it had a plan and how it understood the essence of the position". Deep Blue was no different in essence

Box 1 Artificial intelligence

Artificial intelligence (AI) received its name at a workshop¹⁰ held by John McCarthy at Dartmouth College in New Hampshire in 1956. Marvin Minsky, Allen Newell and Herb Simon, together with John McCarthy, set the research agenda for machine intelligence for the next 30 years. All were inspired by earlier work by Alan Turing, Claude Shannon and Norbert Weiner on tree search for playing chess. From this workshop, tree search — for game playing, for proving theorems, for reasoning, for perceptual processes such as vision and speech and for learning — became the dominant mode of thought, and continues to be so today for large parts of the academic enterprise called AI. Search algorithms are mostly intrinsically serial, which contrasts with the ways in which large parts of animal nervous systems work, where there are both feedforward and feedback signals, but over very low-diameter networks. The speed of the individual neurons is so slow compared with the speed of overall action of the creature that very different intrinsic computations must be at play. There have been numerous flirtations with neurally inspired networks, but most of these are used as component-learning black boxes within a more traditional framework of serial computation.

Game playing is perhaps the most visible success associated with search. Games from tic-tac-toe through to chess have been conquered by brute-force search. More elaborate models of how humans play games have been tried, but in each case performance has soon been overtaken by the unstoppable march of Moore's law. The game of Go, however, has so many possible moves that it has remained impervious to brute-force search. Whereas it seems unlikely that humans play chess by brute-force search, and investigation of methods of play other than search are attractive, Go enforces these investigations with a vengeance. Even when we have computers with the same level of processing power as the human brain, they will not be able to play a good game of Go using brute-force search alone.

Search, however, has also been successful in other areas. It is used in proving theorems, in mathematical manipulation systems such as Mathematica and Matlab, and in systems of speech understanding and natural language understanding.

from the earlier versions he had been playing in the late 1980s. Deep Blue still had no strategic planning phase, as other chess programs designed to model human playing had. It still had only a tactical search, albeit a very deep, fast tactical search. This appeared to Kasparov to be about game plans, not because there was anything new, but because more computer power made the approach

feel qualitatively different. The same might happen to our models of intelligence and life, if we could only get enough computer power.

If any of the above is the case then we should expect great progress in AI and Alife as soon as someone stumbles across the things that need to be fixed. The details will not particularly surprise anyone, although the new developments will have great practical impact. They will lead to new insights in all the sciences that study living organisms, as they will give us new sorts of computer models with which we can test rafts of new hypotheses about how living systems operate.

Models lack unimagined features

But what if we are missing something fundamental and currently unimagined in our models? We would then need to find new ways of thinking about living systems to make any progress, and this will be much more disruptive to all biology. As an analogy, suppose we were building physical simulations of elastic objects falling and colliding. If we did not quite understand physics, we might leave out mass as a specifiable attribute of the objects. Their falling behaviour would at first seem correct, but as soon as we started to look at collisions we would notice that the physical world was not being modelled correctly.

So what might be the nature of this unimagined feature of life? One possibility is that some aspect of living systems is invisible to us right now. The current scientific view of living things is that they are machines whose components are biomolecules. It is not completely impossible that we might discover some new properties of biomolecules or some new ingredient. One might imagine something on a par with the discovery of X-rays a century ago, which eventually led to our still-evolving understanding of quantum mechanics. Relativity was the other such discovery of the twentieth century, and had a similarly disruptive impact on the basic understanding of physics. Some similar discovery might rock our understanding of the basis of living systems.

New stuff

Let us call this the 'new stuff' hypothesis — the hypothesis that there may be some extra sort of 'stuff' in living systems outside our current scientific understanding. Roger Penrose⁸, for one, has already hypothesized a weak form of 'new stuff' as an explanation for consciousness. He suggests that quantum effects in the microtubules of nerve cells might be the locus of consciousness at the level of the individual cell, which combines in bigger wave functions at the organism level. Penrose has not worked out a real theory of how this might work. Rather, he has suggested that this may be a critical element that will need to be incorporated in a final

understanding. This is a weak form of new stuff because it does not rely on anything outside the realm of current physics. For some it may have a certain appeal in that it unifies a great discovery in physics with a great question in biology — the nature of consciousness. David Chalmers⁹ has hypothesized a stronger form of new stuff as an alternative explanation for consciousness. He suggests that a fundamentally new type, of the order of importance of spin or charm in particle physics, say, may be necessary to explain consciousness. It would be a new sort of physical property of things in the Universe, subject to physical laws that we just do not yet understand. Other philosophers, both natural and religious, might hypothesize some more ineffable entity such as a soul or *elan vital* — the 'vital force'.

Another way that the unimaginable discovery might come about is through 'new mathematics'. This would not require any new physics to be present in living systems. We may simply not be seeing some fundamental mathematical description of what is going on in living systems and so be leaving it out of our AI and Alife models. What might this 'new mathematics' be? Candidates have included catastrophe theory, chaos theory, dynamical systems and wavelets. When each of these new mathematical techniques hit the market, researchers noticed ways in which they could be used to describe what is going on in living systems, and then tried to incorporate the same thing into their computational models. It is not clear whether the mathematical techniques in question are best used as descriptive tools or as generative components within the computational models. The latter approach seems at times misguided. However, none of these wonder techniques has really made the hoped-for improvements in our models. Looking at the physical nature of living systems, there seem to be certain mathematical properties that are not handled at all by any of these new techniques, or by any current model. One property is that the matter that makes up living systems obeys the laws of physics in ways that are expensive to simulate computationally. For instance, the membranes of cells have a shape determined by the continuous minimization of forces between molecules within the membrane and on either side of it. Another property is that matter does not simply appear and disappear in the physical world, but great care must be taken in a computational simulation to enforce this.

An analogy to the sort of thing that might be missing is computation — not as the undiscovered feature itself but as an analogy for the type of thing we might be looking for. For most of the twentieth century we have poked electrodes into living nervous systems and looked for correlations between the

Box 2 Artificial life

Chris Langton instigated the notion of artificial life (Alife) at a workshop¹¹ in Los Alamos, New Mexico, in 1987. The enterprise was to make living systems without the direct aid of biological structures. The work was inspired largely by John Von Neumann, and his early work on self-reproducing machines in cellular automata. Von Neumann's model of computation was a Turing machine, the same intrinsically serial mechanism adopted by AI. Researchers in Alife were much inspired by biological systems and quickly produced simulations and analytical models of aspects of biological reproduction and group behaviours of organisms.

Tom Ray¹² developed a system called Tierra, where multiple computer programs competed for the resource of the processing unit in a simulated computer. The seed program could reproduce itself, and took 80 nibbles (half-bytes) of code space. It followed the classical Von Neumann model, and had the central feature of molecular biology, in that the code needed to be both transcribed, or interpreted, and copied into the progeny so that it too had its own 'DNA'. The simulated computer, however, was subject to copying mistakes, and to 'cosmic rays' which randomly flipped bits in the memory space. Its memory filled up with copies of the original seed program, although some of them had errors and were removed. But others started to optimize and get smaller — there was environmental pressure selecting for small programs — and parasites less than half the size of the original programs soon evolved. Although not able to copy themselves, they could trick a larger program into copying them rather than itself. Diverse populations, such as one sees in nature, soon arose in the system.

Most recently, Lipson and Pollack¹³ have evolved 'creatures' that can move. Their morphology is restricted to fixed-length bars with ball-and-socket joints. The designs evolved in a physical simulation program, and the fitness of individuals was determined by measuring their ability to move on a surface. Mutating the most successful members of the previous generation produced each new generation. The *tour de force* was to have the creatures automatically produced directly from the evolved genome by a computerized rapid prototyping system. Motors were snapped on by hand and the creatures, connected to a computer running the same neural network used in the simulations, then started crawling across physical surfaces.

signals measured and events that occur elsewhere in the creature. These data are used to test hypotheses about how the living system 'computes' in the broadest sense of the word. Imagine a society isolated for the past hundred years and in which computers have not been invented. If the scientists in this society came across a working computer, would they be able to understand what it was

doing if they had no notion of computation? Would it make any sense without the notion of Turing computability, or an understanding of a Von Neumann architecture? Or would our isolated scientists need to reinvent the notion of computation before they could explain what the machine was doing? I strongly suspect that they would. Nothing that Turing or Von Neumann did in their mathematics at this level was particularly disruptive. A good late-nineteenth-century mathematician could understand it all with a few days instruction — there would be no surprises for them in the way that quantum mechanics and relativity would surprise a physicist from the same era.

So now we return to the unimaginable. For perceptual systems, say, there might be some organizing principle, some mathematical notion that we need in order to understand how they really work. If so, discovering this principle will enable us to build computer-vision systems that are good at separating objects from the background, understanding facial expression, discriminating the living from the non-living and general object recognition. None of our current vision systems can do much at all in any of these areas. What form might this mathematical notion take? It need not be disruptive of our current view of living things, but could be as non-threatening as the notion of computation, just different to anything anyone has currently thought of. Perhaps other mathematical principles or notions, necessary to build good explanations of the details of evolution, cognition, consciousness or learning, will be discovered or invented and let those subfields of AI and Alife flower. Or perhaps there will be just one mathematical notion, one 'new mathematics' idea, that will unify all these fields, revolutionize many aspects of research involving living systems, and enable rapid progress in AI and Alife. That would be surprising, delightful and exciting. And of course whether or not this will happen is totally unforeseeable.

Rodney Brooks is at the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 545 Technology Square, Cambridge, Massachusetts 02139, USA.

1. Greenblatt, R. D., Eastlake, D. E. & Crocker, S. D. in *Proc. Fall Joint Comput. Conf.* 801–810 (AFIPS Press, Montvale, NJ, 1967).
2. Marr, D. C. & Hildreth, E. *Proc. R. Soc. Lond. B* **207**, 187–217 (1980).
3. Pomerleau, D. A. & Jochem, T. M. *Photon. Spectra* 80–85 (April 1996).
4. Kitano, H. *New Generation Comput.* **18**, 199–216 (2000).
5. Brooks, R. A. *Science* **253**, 1227–1232 (1991).
6. Adams, B., Breazeal, C., Brooks, R. A. & Scassellati, B. *IEEE Intell. Syst. Applic.* **15**, 25–31 (2000).
7. Matijevic, J. *Science* **280**, 454–455 (1998).
8. Penrose, R. *Shadows of the Mind* (Oxford Univ. Press, New York, 1994).
9. Chalmers, D. J. *Conscious. Stud.* **2**, 200–219 (1995).
10. McCorduck, P. *Machines Who Think* (Freeman, New York, 1979).
11. Langton, C. G. (ed.) *Artificial Life* (Addison-Wesley, CA, 1989).
12. Ray, T. S. in *Artificial Life Vol. II* (ed. Langton, C. G.) 371–408 (Addison-Wesley, CA, 1991).
13. Lipson H. & Pollack, J. B. *Nature* **406**, 974–978 (2000).

Life's lessons in design

Philip Ball

So long as it avoids a Panglossian view of nature, the science of biomimetics has the potential to enrich many areas of technology. But accurate mimicry will require greater understanding of natural mechanisms at the molecular scale. As this continues to unfold, emulation may increasingly give way to assimilation of biological machinery.

There is such a long and colourful history of engineers, scientists and artificers gaining inspiration from nature that one could be forgiven for thinking that all the best ideas have been spoken for. In the nineteenth century, biomimesis was at least as much an aesthetic as a practical pursuit. Artists and architects delighted in Ernst Haeckel's drawings of radiolarians for their beauty alone. When the French designer René Binet conceived of the elaborate entrance gate to the World Exposition in Paris in 1900, he told Haeckel: "everything about it, from the general composition to the smallest details, has been inspired by your studies"¹ (Fig. 1).

But others recognized the inventiveness, economy and sound engineering of nature's structures. The Wright brothers took flight after watching vultures swoop, giving a nod to Leonardo da Vinci's explicitly aviamorphic flying machines. Joseph Paxton is said to have paid tribute to the ribbed stem of a lily leaf in his Crystal Palace, which housed the Great Exhibition of 1851. Gustave Eiffel's tower supports its own immense weight along elegant curves inspired by bone structure. D'Arcy Thompson² tells how in 1866 the engineer C. Culmann in Zürich, pondering on the design of a new construction crane, wandered into the laboratory of the anatomist Hermann Meyer who was studying cross-sections of bone. Observing how the trabeculae of the porous material traced out lines of tension and compression, he cried out: "That's my crane!"

This rich heritage means that the diverse array of scientists and technologists who today take their lead from nature may feel they are immersed in the paradigm of an earlier age, with its traditional-sounding considerations of morphology, stress distribution and hydrodynamic forces. And yet the materials and devices emerging from biomimetics are unmistakably forward-looking: new solar cells, smart sensors, advanced robotics and aerospace materials.

But today, biomimetics has something much more dramatic to offer than an aircraft wing or an anti-drag surface coating modelled after some natural example. One of the biggest obstacles to taking full advantage of what nature has to offer is that the living

world has an awesomely elaborate means of construction. There is no assembly plant so delicate, versatile and adaptive as the cell. But as modern methods of investigation and analysis decode and elucidate the cell's molecular machinery piece by piece, this disparity between the natural and synthetic art of manufacture begins to diminish. When biomimicry proceeds at the molecular scale, as it is now beginning to do, its entire basis is transformed.

And the inevitable corollary of molecular biomimetics is a new conceit, which one might call biosynergic engineering: merging nature's machinery with synthetic constructs to develop a new kind of synthetic methodology at the molecular scale. The scattered, primitive beginnings of such a movement already reveal that nature may have even more potential than it displays in the wild.

Small is beautiful

Biomimetics has encompassed some grand engineering, but it is undoubtedly small science — indeed, often budget science. The reasons are clear: nature does not employ exotic materials, nor extreme energies or high pressures. The living world comes almost with

a guarantee of economy, for that is evolution's exigency. Maximum return for minimal (metabolic) outlay: this is the stipulation that keeps nature lean and, in some sense, optimal.

Yet in what sense, exactly? The engineer must consider that carefully. It would be foolish to assume that natural selection stands proxy for the testing laboratory or the market, refining a design in just those ways that a new product demands. The complex of compromises that shape the fitness landscape of evolution is likely to bear only incidental correspondences with that which determines the contours of technological and economic viability. This, after all, is why we seek to mimic and not to duplicate. One of the attractions, as well as one of the main challenges, of biomimetics is that it demands creative solutions. Nature's pool of ideas is valuable only if it can be translated into terms that the technologist can work with, particularly in terms of materials and processing methods.

Take wood, for example. There has been relatively little serious attempt to produce an artificial analogue, for the simple reason that wood itself is already an almost peerless structural material for certain applications:

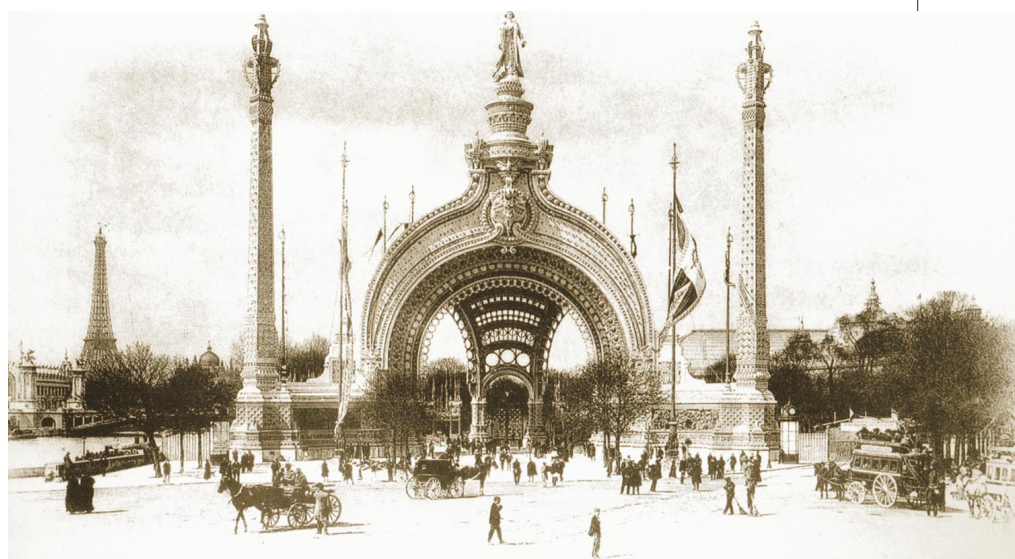


Figure 1 René Binet's entrance to the World Exposition in Paris, 1900, inspired by Haeckel's drawings of radiolarians.

cheap, lightweight, tough, mouldable and easily shaped. But it is not perfect, especially in terms of durability in the face of damp and pests. Yet which features of wood's enormously complex structure are the most salient for mimicry in a synthetic version?

The material developed 20 years ago by Gordon and Jeronimidis³ latches onto the fibrous structure to capture fracture-resistance, and so uses glass fibre in a resin matrix. The low-density cellular structure is only crudely imitated in this material, by corrugating some of the laminated layers — this suffices to keep the material light without sacrificing too much robustness.

But there is more to be learned from the natural engineering of wood. Many engineering materials must be punctured for joining purposes. But it is one of the oldest principles in the engineer's handbook, notorious ever since Inglis worried nearly a hundred years ago why British ships were breaking in half, that holes produce stress concentration and so allow cracks to nucleate. The fibres that reinforce wood's glassy lignin matrix are severed and rendered structurally ineffective where a hole is drilled in timber.

The tree, however, drills no holes, even though it must disrupt the trunk's wood where a new branch pushes through. The solution is obvious to see in planking: the fibres deform around a knothole, remaining continuous. This simple solution avoids a significant reduction in fracture strength, yet has been little exploited in fibrous composite materials. Jeronimidis is now proposing to do so⁴.

High fracture strength is also the alluring aspect of the abalone shell nacre, in many ways the type specimen for biomimetic materials science. Nacre combines several of the features recognized as characteristic of how the properties of superior materials are engineered in natural substances. It is a composite material, and incorporates both inorganic (here calcium carbonate) and organic compounds. The microscopic structure is finely wrought: plates of the hard mineral interleave with sheets of proteins and other macromolecules. The mineralization process is highly controlled, promoting the less thermodynamically stable crystal polymorph (aragonite rather than calcite) with crystallographic planes aligned in different platelets and the crystal morphology constrained to flat sheets. Some of these features recur in bone, eggshell, tooth enamel and the exoskeletons of diatoms and radiolarians.

The macromolecules are presumed responsible for guiding nucleation and growth of the mineral, but beyond that much is mysterious. The idea that acidic groups in the protein sheets provide a template for the aragonite crystal planes now looks overly simplistic; soluble acidic proteins instead seem to be a major determinant of crystallization⁵. But the growth process is not just a

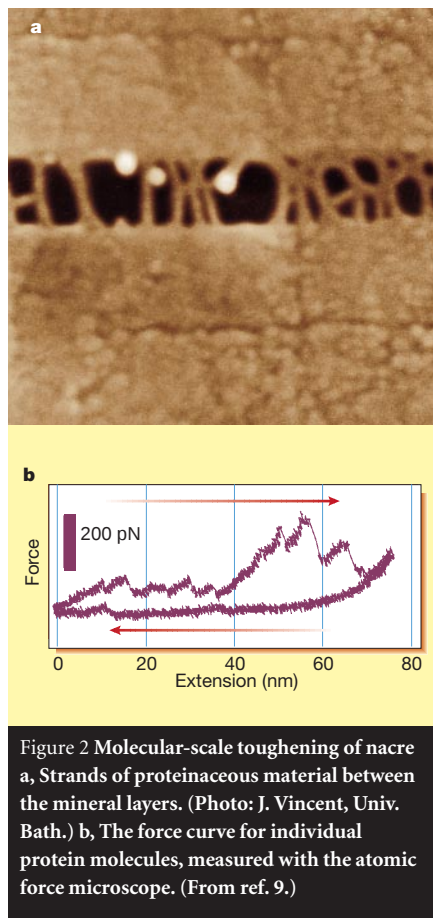


Figure 2 Molecular-scale toughening of nacre **a**, Strands of proteinaceous material between the mineral layers. (Photo: J. Vincent, Univ. Bath.) **b**, The force curve for individual protein molecules, measured with the atomic force microscope. (From ref. 9.)

matter of engineering growth in the places and shapes it is required; it also demands that nucleation is suppressed elsewhere in a fluid that is supersaturated with the salt.

It seems quite possible that the soluble proteins act in a manner comparable to that of the antifreeze proteins in cold-water fish, which can both inhibit and promote ice nucleation according to taste. These proteins are thought to have repetitive hydrogen-bonding groups commensurate with the lattice spacings in ice.

Understanding the molecular-scale mechanisms of the nucleation and growth (two distinct phenomena, don't forget) of crystals promises dividends in the industrial preparation of metastable polymorphs, or the suppression of degradative mineralization on engineering structures. *De novo* design of crystallization regulators is still at an early, exploratory and largely empirical stage⁶, although combinatorial methods including immunization techniques for raising crystal-binding antibodies⁷ look encouraging.

The toughening mechanism of nacre — crack deflection and energy absorption at 'weak' interfaces — is now well established, and has been demonstrated in synthetic laminated materials. But close inspection shows that there is more to it than that, reminding us that nature is cautious and generally seeks several simultaneous solutions to a challenge. As the mineral plates are pulled

apart during deformation and fracture of the shell, tiny strands are pulled out from the intervening organic layers⁸ (Fig. 2a). Single-molecule measurements of the force-extension curve of these strands with the atomic force microscope show that they lengthen in 'modular' fashion, producing a series of sawtooth jumps⁹ (Fig. 2b). By combining relative stiffness (before each jump) with large extensibility, this creates a high work of fracture (equal to the area under the force curve). The same principle seems to be at work in the titin molecule, the elastic cord that prevents the interdigitating sarcomeres from separating when skeletal muscle is highly extended. Here the sawtooth pattern has been shown to have a clever structural origin: the protein consists of a series of identical globular domains that unravel one by one¹⁰.

Ant strategies

To appreciate that nature does not necessarily have all the best ideas, we need only point to the wheel. Nevertheless, most of the problems of controlled motion and manoeuvrability have been explored and elegantly resolved in the living world. The Wrights and Otto Lilienthal drew comfort from the evident fact that they were not attempting the impossible.

Yet nature shows too how all flight is not the same. The smaller an airborne creature gets, the more manoeuvrable it typically is, which tells the engineer at once about the importance of Reynolds number: in a medium of fixed viscosity, aerodynamic phenomena have a particular size scale. It is becoming gradually clear that this allows insects access to different tricks, and thus different flight patterns, than buzzards and gulls.

Specifically, insects are conjurors of the vortex. With deft flappings and rotations of their wings, they are able to manipulate the vortices shed from the edges to control their motion in ways that flight engineers can only dream of: taking off backwards, for example, or landing upside down. By such means, insects subvert the 'conventional' aerofoil principles of flight, giving rise to the canard that the bee is aerodynamically impossible. In essence, the flight of the bumble-bee is a flight beyond the dynamical steady state: lift is generated at particular, exquisitely timed moments during the flap cycle. By rotating the wing so that it is parallel to the ground on the downstroke but perpendicular on the recovery stroke, an insect is able to recapture energy from the vortices shed from the wing edge¹¹. This reveals a new mechanism for flight that one could hardly have deduced from first principles, and which might be adopted for the development of miniaturized robotic flyers for remote sensing, surveying and planetary exploration.

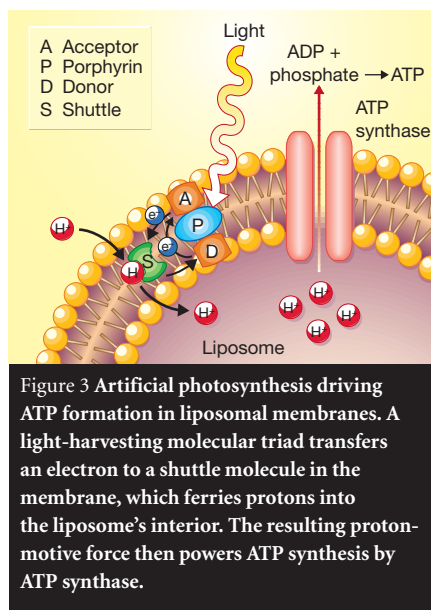
The actuators and sensors needed to drive and direct robot motion can benefit from studies of nature. Here an important part of the motivation may be simplicity. There has

been a tendency to overdesign robots, imbuing them with actuators that permit every conceivable mechanical rearrangement of limbs and with sensors that provide exhaustive information about the surroundings. Nature, always seeking economy, makes do with far less. An understanding of how animals navigate — by landmarks, trail laying, geomagnetism or mental integration of the path already covered — should indicate the minimal requirements in different landscapes. And the articulation of moving parts often provides a lesson in creative simplification. Instead of having active control mechanisms for movement in each direction, limbs are often moved by the passive properties of the hinge materials. In insects, for example, the wing hinge is not attached to the flight muscles at all. Rather, these muscles deform the elastic, resilient cuticle of the thorax, which translates its change of shape to a wing oscillation¹². This mechanism, which permits small strains to give rise to large-amplitude motion, is surely instructive to engineers wondering how to extract large displacement from the generally small strains available from piezoelectric smart actuator materials. In fact, a similar principle is already used in the crescent-shaped 'moonie' actuators devised by Newnham and co-workers¹³, and in the piezoelectric heads in some dot-matrix printers.

One of the most striking messages for robotics coming from the study of animal motion is that some tasks are more efficiently conducted by many small, simple entities than by a single large and complex one. Approaches to robot design that attempt to search 'design space' in a pseudo-evolutionary way rather than to impose a preconceived strategy¹⁴ may now include the option of fragmentation: of allowing for a distributed solution¹⁵. This is, of course, how an ant colony operates as a kind of 'super-organism' when foraging.

A recent study of ant-mimicking division of labour in a swarm of 'cooperative' robots lacking decentralized control showed that this does indeed provide an efficient foraging strategy¹⁶. The robotic swarms exhibit an optimal size, reminiscent of the size selection seen in animal colonies: for larger groups, the (programmed) tendency to avoid other robots begins to inhibit a thorough search of the territory. Social insects often operate task recruitment, whereby one communicates with another to draw it towards a resource-rich area. The inclusion of such a capability in the robot swarm boosts its ability to exploit clustered resources.

Search algorithms derived from those of ant colonies confer benefits not only in real-space tasks, but also in computation. Of course, the original 'biomimetic' computer application is the genetic algorithm. But some search tasks, such as trawling a large



database (as in a web search engine), lend themselves to schemes that implement a kind of trail-laying, like the pheromones deposited by foraging ants¹⁷. 'Evaporation' of the trail is essential to prevent chasing of false leads or movement towards depleted sources, and the efficiency of the search can be highly dependent on identifying the optimal evaporation rate.

Search for systems

Engineers interested in flight, or in tough materials, do not find it hard to know where to look in nature (even if the answers they find are subtle and hard to tease out). But the routes of technology transfer from biology to engineering sciences are not always so obvious. There might be a myriad of good ideas buried in the living world for an engineer searching for innovation in a particular area — but how to find them? "In nature there are lots of hidden patents," is the rather mercenary but nevertheless apt way that the Russian engineer Genrich Altshuller expresses it.

Altshuller has attempted the apparently oxymoronic task of systematizing creativity in technical innovation. By analyzing over a million engineering patents, he has identified 39 'principles' on which particular engineering problems hinge. The problems then arise from situations in which one of these principles or parameters has to satisfy conflicting requirements. The 'contradiction matrix' thus contains 1,482 (39 × 38) 'standard technical conflicts', which Altshuller suggests can be addressed by a series of 'standard solutions'. He calls this the Theory of Inventive Problem Solving, denoted by its Russian acronym TRIZ¹⁸.

Altshuller claims that almost all innovation requires knowledge already available, if sometimes from disciplines far removed from the immediate one. TRIZ claims to ease that process of information retrieval. Julian

Vincent has considered how TRIZ fares if applied to the engineering problems faced by living organisms¹⁹. The assessment, while anecdotal, is revealing.

First, it is possible to identify natural analogues of all 40 of the standard solutions, which include such things as segmentation of parts, asymmetrical design, nesting of objects, multiple functionality and porosity. This in itself implies some kind of mapping between biology and technology. Yet nature does not seem to use the same contradiction matrix to solve problems²⁰. For example, the drag-reducing properties of shark skin make it one of the archetypal examples of natural engineering. But identifying on the TRIZ matrix the contradictions that arise in this hydrodynamic problem suggests that appropriate solutions should involve weight compensation, moving parts or changes in some ambient parameter such as compliance. In contrast, shark skin uses a solution not included in Altshuller's list: surface conformation. The microscopic ribs of the shark's scales suppress turbulence in the boundary-layer flow.

This does not imply that TRIZ is a useless concept. Rather, it suggests that human ingenuity is a restricted resource: several thousand good engineering heads cannot compete with the billions of years that evolution has had to experiment, select and refine. In other words, says Vincent, nature may be a treasure trove for that small but vital fraction of engineering problems that, in Altshuller's reckoning, require genuine innovation and even fresh discovery.

Biosynergy

Yet the question remains: are nature's solutions practically accessible to us? It might not be fruitful, or even possible, to isolate and imitate one aspect of a functioning organism while neglecting the dynamic system in which it is embedded, and which is responsible for its fabrication, maintenance and adaptation. Silk is a sobering example.

No one expected that simply mimicking the amide linkages in the polymer chains, as in nylon, would generate a material quite as appealing. But Kevlar, the aramid fibre in which intermolecular hydrogen bonding creates a degree of liquid crystallinity similar to that in concentrated silk solution, deepens the mimesis and greatly improves the strength. Shear-induced alignment of polyethylene during the extrusion process, copying that which occurs in the silk spider's spinneret, produces the high-strength fibre known as 'rocket wire'.

Best of all, one might think, is to take mimicry to the point of plagiarism, and copy the silk protein exactly. But although silks have been sequenced and silk genes spliced into the bacterium *Escherichia coli* and goats (which express the protein in their milk), synthetic silk is still not a mass-produced, high-strength technological material. The

crucial aspect of mimicry is not, it seems, in the protein composition but in the processing. It is the weaving of strands in the spinneret that gives them their strength. The details of this process are not understood; but it may be that not until we can build an artificial, miniaturized spinning mechanism will silk be an industrial material.

This is why biomimetics must reach down to the microscopic and ultimately the molecular scale. Some of nature's best tricks are conceptually simple and easy to rationalize in physical or engineering terms; but realizing them requires machinery of exquisite delicacy. "The smaller the scale [at which mimicry is conducted], the better the prospects for emulation," says Steven Vogel, who points out that nature's artefacts are made in factories smaller than their products²¹. It is precisely this 'bottom-up' approach to fabrication that is being sought within the field of nanotechnology — which, ever since its beginnings in Richard Feynman's famous talk²², has acknowledged the inspiration and guidance that biology offers.

Photosynthesis, for instance, is a trick worth mastering. This is not a matter of efficiency: commercial silicon solar cells already do several per cent better at converting light to electrical energy than the chloroplast does in making the conversion to chemical energy. But the nanocrystal solar cells developed by Grätzel and co-workers²³ show the benefit of the chloroplast's design principles. Allocating charge generation and charge separation to different entities — in the leaf, to chlorophyll and to pheophytin, plastoquinone and other elements of the electron relay — reduces the chances of recombination of the light-excited hole and electron. In silicon solar cells, the semiconductor serves both ends, and recombination can limit the quantum efficiency of the process. Grätzel's concept was to capture the photon energy using dye molecules adsorbed to the surface of nanocrystals of titania. These semiconducting particles then ferry the charge to the collecting electrode. This helps to secure a respectable efficiency that, although by no means outstanding in itself, promises a competitive device when coupled to the very low manufacturing and materials costs.

But more transparently biomimetic is the liposome-based system developed by Moore, Gust, and their co-workers^{24,25}. This is speculative mimicry — emulation without a current application — for the sunlight is here used to make ATP, as it is in the chloroplast, rather than to generate a flow of current. The motivation might therefore be construed as more akin to biocatalysis than photovoltaics: storing up photonic energy in a form accessible to biological systems. The liposomes mimic the thylakoid membrane of the chloroplast, anchoring and organizing the light-harvesting and energy-transfer molecules while also providing a barrier across

which an electrochemical gradient can be established. The role of the photosynthetic reaction centre and antenna array is adopted by a synthetic molecular triad in which a photoexcited porphyrin passes energy to an electron donor, which releases an electron to an acceptor group. From here it is transferred to a mobile 'shuttle' molecule within the membrane, the surrogate for the electron carrier NADPH in photosynthesis. The electron-charged shuttle picks up a hydrogen ion too and carries them both to the inner face of the membrane. There it returns the electron to the triad, and releases the hydrogen ion into the liposome's hollow interior. This light-driven hydrogen-ion pump thus creates a proton-motive force, which is harnessed by ATP synthase embedded in the membrane (Fig. 3).

As a demonstration that a complex natural molecular process can be imitated in a self-assembling synthetic system, this work carries an encouraging message to deter any vitalistic suggestion that life's mechanics are incomparably intricate. But perhaps it stretches the definition too far to regard as genuinely biomimetic a system that uses molecules such as ATP synthase ready-made?

Of course, it is eminently sensible to do so, rather than trying to devise a proton-driven catalyst *de novo*. There is thus good reason to imagine that any molecular-scale engineering that seeks to achieve things at which the cell is already adept — such as energy conversion, construction or replication — will be wise to incorporate, rather than to imitate, biological machinery.

This sort of biosynergy has been elegantly explored with motor proteins. Work on wholly synthetic molecular motors is still in its infancy, and so far takes few cues from nature^{26,27}. But the possibility of modifying motor proteins to do non-natural tasks is already apparent^{28,29}. For example, Dennis *et al.*²⁹ have used immobilized kinesins oriented along corrugations in shear-aligned polytetrafluoroethylene films to transport microtubules across a surface with directional preferences. The development of peptides by *in vitro* selection that can recognize different metal³⁰ and semiconductor³¹ surfaces suggests the possibility of using genetic recombination methods to append these selective hooks to motor proteins to transport and organize semiconductor nanoparticles (quantum dots) into arrays for information technology. In any event, molecular nanotechnologists would surely be short-sighted if they did not merely take inspiration but also working machines and devices from the cell.

Choices and challenges

Vogel²¹ presents elegant and persuasive arguments for why it would be foolish to assume that nature has all the best ideas, which the engineer must then determine how to translate into workable solutions. The caricature of evolution in which nature

explores all options and finds the best is still surprisingly pervasive. Nature has good reasons to avoid metallic components, for example, but this does not mean that human engineers should strive to do so.

Yet fundamental research on the character of nature's mechanisms, from the elephant to the protein, is sure to enrich the pool from which designers and engineers can draw ideas. The scope for deepening this pool is still tremendous. It is at the molecular scale, however, that we will surely see the greatest expansion of horizons, as structural studies and single-molecule experiments reveal the mechanics of biomolecules. If any reminder were still needed that nanotechnology should not seek to shrink mechanical engineering, cogs and all, to the molecular scale, it is found here. Nature's wheel — the rotary motor of the bacterial flagellum — never got any larger than this, nor is it fashioned from hard, wear-resistant materials, nor is it driven electromagnetically or by displacement of a piston. But it is efficient, fast, linear and reversible³². Somewhere there is a lesson in that.

Philip Ball is a consultant editor for Nature.

- Krause, E. in *Die Rezeption von Evolutionstheorien im 19. Jahrhundert* (ed. Engels, E.-M.) (Frankfurt, 1995).
- Thompson, D. W. *On Growth and Form* (Cambridge Univ. Press, 1942).
- Gordon, J. E. & Jeronimidis, G. *Phil. Trans. R. Soc. Lond. A* **294**, 545–550 (1980).
- Jeronimidis, G. *Proc. Workshop on Bionics and Biomimetics* (Wissenschaftskolleg, Berlin, 1998).
- Belcher, A. M. *et al. Nature* **381**, 56–58 (1996).
- Coveney, P. V. *et al. J. Am. Chem. Soc.* **122**, 11557–11558 (2000).
- Bromberg, R., Kessler, N. & Addadi, L. *J. Cryst. Growth* **193**, 656–664 (1998).
- Jackson, A. P., Vincent, J. F. V. & Turner, R. M. *Proc. R. Soc. Lond. B* **234**, 415–440 (1988).
- Smith, B. L. *et al. Nature* **399**, 761–763 (1999).
- Fisher, T. E., Marszalek, P. E. & Fernandez, J. M. *Nature Struct. Biol.* **7**, 719 (2000).
- Dickinson, M. H., Lehmann, F.-O. & Sane, S. P. *Science* **284**, 1954–1960 (1999).
- Dickinson, M. H. & Tu, M. S. *Comp. Biochem. Physiol. A Physiol.* **116**, 223–238 (1997).
- Trolier-McKinstry, S. & Newnham, R. E. *MRS Bull.* April, 27–33 (1993).
- Lipson, H. & Pollack, J. B. *Nature* **406**, 974–978 (2000).
- Kube, C. R. & Zhang, H. *Adapt. Behav.* **2**, 189–218 (1994).
- Krieger, M. J. B., Billeter, J.-B. & Keller, L. *Nature* **406**, 992–995 (2000).
- Bonabeau, E., Dorigo, M. & Theraulaz, G. *Nature* **406**, 39–42 (2000).
- Altshuller, G. *And Suddenly the Inventor Appeared—TRIZ, the Theory of Inventive Problem Solving* (Technical Innovation Centre, Worcester, MA, 1999).
- Vincent, J. F. V. <<http://www.triz-journal.com/archives/>> *TRIZ J.* August (2000).
- Mann, D. <<http://www.triz-journal.com/archives/1999/11/c/index.htm>> *TRIZ J.* November (1999).
- Vogel, S. *Cats' Paws and Catapults* (Penguin, London, 1999).
- Feynman, R. *Engineering & Science* February 1960. <<http://www.zyvox.com/nanotech/feynman.html>>.
- O'Regan, B. & Grätzel, M. *Nature* **353**, 737–740 (1991).
- Steinberg-Yfrach, G. *et al. Nature* **385**, 239–241 (1997).
- Steinberg-Yfrach, G. *et al. Nature* **392**, 479–482 (1998).
- Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. *Nature* **401**, 152–155 (1999).
- Kelly, T. R., De Silva, H. & Silva, R. A. *Nature* **401**, 150–152 (1999).
- Nédélec, F. J., Surrey, T., Maggs, A. C. & Leibler, S. *Nature* **389**, 305–308 (1997).
- Dennis, J., Howard, J. & Vogel, V. *Nanotechnology* **10**, 232–236 (1999).
- Brown, S. *Nature Biotechnol.* **15**, 269–272 (1997).
- Whaley, S. R., English, D. S., Hu, E. L., Barbara, P. F. & Belcher, A. M. *Nature* **405**, 665–668 (2000).
- Fung, D. C. & Berg, H. C. *Nature* **375**, 809–812 (1995).

Earth systems engineering and management

Stephen H. Schneider

Imagine that we could let the world's economy continue to grow, bring the disadvantaged classes up from poverty and at the same time not threaten the atmosphere or global ecosystems with unprecedented build-up of greenhouse gases and the projected climatic risks of such growth. Earth systems engineering and management may just be such a panacea, some have suggested. But could we anticipate the costs or ever truly predict the consequences?

Few people would ask us to accept that a growing world economy based on greatly expanded per capita energy consumption would be free of environmental side effects. But many have claimed that the anticipated several-fold increase in greenhouse gases — and associated sea-level rises, intensified hurricanes, or drought and flood stresses — can be largely overcome by human ingenuity. Their optimistic vision depends greatly on what had been called 'geoengineering' and has more recently been relabelled Earth systems engineering. This describes the deliberate manipulation of the Earth system to manage the climatic consequences of human population and economic expansion¹.

To others, the notion of geoengineering — injecting dust in the stratosphere, for example, to reflect some sunlight back to space and counteract greenhouse warming — is an irresponsible palliative. It evades the need for a real cure, such as curbing the consumption of the rich and the population growth of the poor, and charging polluters for their use of the atmosphere as a free sewer.

In response, defenders of geoengineering retort that two-thirds of the world's people use a small fraction of the energy per capita of the rich. Cheap primary energy (mainly coal) is needed, they say, to build the economies of less developed countries and improve their well being. The negative environmental side effects of this will have to either be tolerated or be sidestepped by geoengineering in order to have it both ways — a materialistic growth-oriented world and relatively undisturbed climate.

At times this debate takes on an ideological tenor. Claims that the imperative of development cannot be impeded by the prospect of global warming are greeted with the assertion that creating inadvertent damage to nature is bad enough, but deliberately attempting to manipulate the climate just to let our old habits prevail is a violation of stewardship and an ethical transgression

against the natural world. These sets of opposing world views — anthropocentric expansion versus stewardship — are not new. They flared in the 1970s with Club of Rome debates over the 'limits to growth' and matured with the publication of the Brundtland Commissions' middle path, aiming to pursue 'sustainable development'². Today, they continue in arguments over whether nations must meet their emissions reductions agreed in the Kyoto Protocol by domestic cuts — even if not cost effective — or be permitted to shop around to buy their obligations elsewhere in the world at lower costs.

Let us return to the central question of what best characterizes Earth systems engineering. Is it a panacea for sustainable development built with vision and ingenuity or a palliative to avoid fundamental limits and maintain the privileged status quo for special interests? There is no easy answer to this question, but I do believe that both sides have merit in parts of their arguments. Here I will try to sketch out some opportunities and

pitfalls that might help to clarify the role of geoengineering and carbon management strategies in the climate policy debate.

Historical perspective

In Homer's *Odyssey*, Ulysses is the frequent beneficiary (or victim) of deliberate weather modification schemes perpetuated by various gods and goddesses. In Shakespeare's *The Tempest*, Prospero, a mortal (albeit one with magical powers), conjures up a tempest to strand on his mystical island a passing ship's crew. In literature and myth, only gods and magicians could control the elements. But in the twentieth century, serious proposals for the deliberate modification of weather and/or climate came from engineers, futurists or those concerned with counteracting the inadvertent anthropogenic modification of the Earth's climate.

About 1960, Rusin and Flit³ from the former Soviet Union published a long essay entitled *Man versus Climate* in which they suggested 'improving' our planet by, for instance, diverting rivers from the Arctic to the Russian wheat fields, or from the Mediterranean to irrigate areas in Asian USSR. One of their ambitious projects was to create a 'Siberian sea' with water taken from the Caspian Sea and Aral Sea areas. Of course, flowery rhetoric with images of blooming arid zones stands in stark contrast to the ecological disaster that surrounds the Aral Sea today, where environmental degradation is associated with much less radical geoengineering projects⁴.

Other such proposals have become part of geoengineering folklore and include damming the Gulf Stream, the Bering Straits or the Nile, or creating a Mediterranean drain back into central Africa where a 'second Nile' would refill Lake Chad, turning it into the 'Chad Sea' after the Straits of Gibraltar were dammed (Fig. 1). But the potential side effects if these projects misfire are rarely discussed — which is not unlikely, given the complexity of the highly nonlinear climate system.



Figure 1 Some geoengineering projects, such as this plan for the irrigation of the Sahara by creating a 'second Nile' to refill Lake Chad, have become part of geoengineering folklore. (Reproduced from ref. 3.)

In the early 1970s, Russian climatologist Mikhail Budyko⁵ suggested that it was “incumbent on us to develop a plan for climate modification that will maintain existing climatic conditions”. What he endorsed was a stratospheric particle layer to reflect away enough sunlight to counteract greenhouse warming. But, wisely, he added the caveat that deliberate climate modification would be premature before the consequences could be calculated with confidence, a task for which the current simplified theories were inadequate.

William Kellogg and I looked at many such schemes in the 1970s and concluded then⁶ that tampering blindly with the weather system would be the height of irresponsibility. Moreover, it would lead to disputes as any natural weather disaster occurring during deliberate climate modification experiments might well be blamed on the climate modifiers. We offered a modest proposal for ‘no-fault climate-disaster insurance’: if a large segment of the world thought that the benefits of a proposed climate modification scheme would outweigh the risks, they should be willing to compensate those who subsequently lost their favoured climate.

Ironically, perhaps, the term ‘geoengineering’ first seems to have been applied to a scheme that is no longer called by that name. It was informally coined by Cesare Marchetti⁷ who outlined a proposal for tackling the problem of CO₂ in the atmosphere by a kind of ‘fuel cycle’ for fossil fuels. Under this proposal, CO₂ would be collected at certain ‘transformation points’ such as the smokestacks of principal fossil fuel-burning industrial centres. It would be disposed of by injection into sinking thermohaline currents (say, the Mediterranean undercurrent entering the Atlantic at Gibraltar) that would carry and spread it into the deep ocean. Today, this kind of a plan is referred to as industrial carbon sequestration, which is part of ‘carbon management’ — controlling the amount of greenhouse gases in the atmosphere. Geo-engineering as a term has evolved to mean deliberate modifications to biogeochemical or energy flows in the climate system. This kind of tampering with natural processes, not surprisingly, inflames passionate debate.

Since Marchetti’s paper, perhaps the most ambitious attempt to justify and classify a range of geoengineering options was associated with a US National Academy of Sciences (NAS) National Research Council panel on the policy implications of global warming⁸.

As a member of that panel, I can report that the very idea of including a chapter on geoengineering led to serious internal and external debates. Many participants (including myself) were worried that even the thought that we could offset some aspects of inadvertent climate modification by deliberate modification schemes could be used as an excuse to continue polluting. Critics instead favoured market incentives to reduce

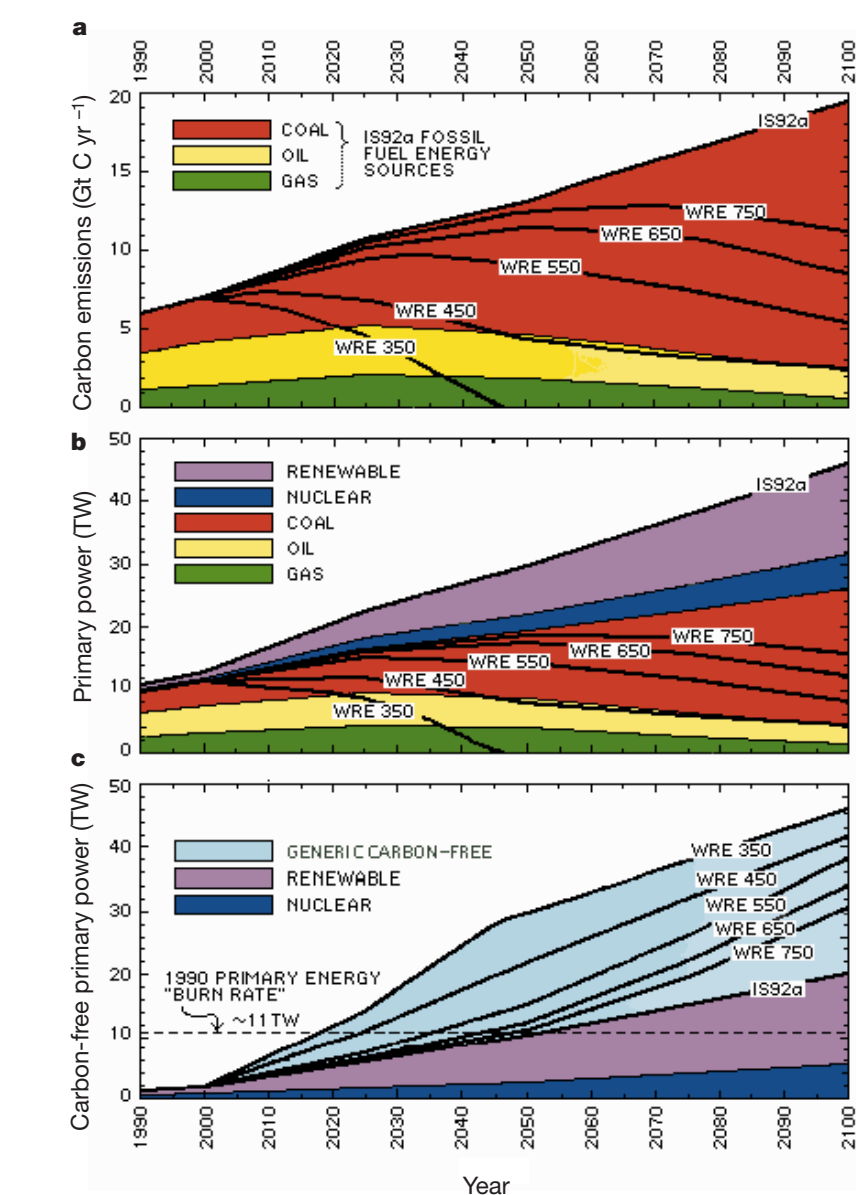


Figure 2 Fossil-fuel carbon emissions and primary power in the twenty-first century for various stabilization scenarios (IPPC scenario IS92a (dubbed ‘business as usual’), and stabilization of atmospheric CO₂ at 750, 650, 550, 450 and 350 p.p.m.v.). a, Carbon emissions; b, primary power; and c, carbon-free primary power (see ref. 13 for further explanation). Carbon-free primary power is total primary power less fossil-fuel carbon power, calculated on the basis of net CO₂ emissions whether from sequestration or solar, nuclear or wind power. (Reproduced from ref. 13.)

emissions or regulations for cleaner alternative technologies. But Robert Frosch countered as follows: what if a pattern of change currently thought unlikely, but of high consequence, actually started to unfold in the decades ahead? It would take decades to develop the technical and political tools to reverse the risks. We would simply have to practise geoengineering as the ‘least evil’.

Although sceptical about the viability of specific engineering proposals and the questionable symbolism of suggesting that we could sidestep real reductions in emissions, I nonetheless voted reluctantly with the majority of the NAS panelists who agreed to allow a carefully worded chapter on the geo-

engineering options to remain in the report.

Extending Budyko’s focus on the injection of aerosol particles (particles suspended in a gas) in the stratosphere, the geoengineering chapter of the report suggested that 16-inch naval rifles fired vertically could propel a 1-ton shell consisting of dust particles up to an altitude of 20 kilometres. Given an aerosol lifetime in the stratosphere of two years, 10 megatons (10¹⁰ kilograms) could be placed in the stratosphere 20 times during a 40-year period until 2030. Over this time the NAS authors estimated geoengineering costs to be about US\$5 per ton carbon (as CO₂) mitigated. This cost is somewhat comparable to carbon taxes proposed by Nordhaus⁹

for modest control of CO₂ emissions. But for a major mitigation of CO₂ emissions (for example, a 20% cut), Nordhaus's study suggests that the carbon taxes required could be hundreds of dollars per ton carbon. (Conventional calculations of the costs of CO₂ mitigation through carbon taxes use economic models that are likely to overestimate the costs of mitigation as these models still ignore the effects of climate policies in inducing technological improvements; for a critique, see ref. 10.)

But is it even possible to inject dust in the stratosphere, for example, in a manner that would perfectly offset a given injection of greenhouse gases in the atmosphere? Even though the 30% increase in CO₂, 150% increase in methane, and the addition of unnatural chemicals such as chlorofluorocarbons have spread fairly uniformly over every square metre of the Earth since the industrial revolution, the patterns of heat trapped as a consequence are not uniform. The primary reason is the non-uniform distribution of other optically active constituents of the atmosphere, especially clouds.

Furthermore, humans add aerosols as well — not primarily the stratospheric kind, but mostly tropospheric sulphate aerosols resulting from the burning of coal and oil. These short-lived, lower-atmospheric aerosols are patchy in distribution and probably reject sunlight back to space at the rate of up to 1 watt per square metre averaged over the Northern Hemisphere¹¹, enough to offset perhaps one-quarter to one-half of the extra infrared heat associated with the enhanced greenhouse effect globally. And biomass burned also produces patchy distributions of aerosols, some of which actually warm the climate since they contain light-absorbing soot, as do some industrial aerosols as well.

Because of the patchy nature of the greenhouse effect itself, even if we could engineer our stratospheric aerosol injections to balance on a hemispheric (or global) basis the amount of hemispherically (or globally) averaged heat trapped by human-contributed greenhouse gases, we would still be left with some regions heated to excess and others left cooler. I am not saying that such anomalies arising from aerosol geoengineering would necessarily be worse than, say, an unabated 5°C warming. But this is why the strong caveats in the NAS report are reiterated by all responsible people who have addressed the question.

As a postscript to this question, a climatic model study at the Lawrence Livermore National Laboratory¹² has actually attempted to simulate whether the zonal patterns of stratospheric aerosol cooling could offset the more patchy patterns of greenhouse gas heating. They concluded optimistically that within the sampling precision of the model — which is still quite noisy — the aerosol scheme might not generate major regional

climatic anomalies relative to those of unabated climatic change. Although not definitive, such studies are needed to give confidence in the effectiveness of any geoengineering scheme. And without high confidence in the outcome, any implementation would be controversial or indeed lead to overt conflicts — the subject we turn to next.

Caretakers for a century?

No institutions currently have the authority to enforce responsible use of the global commons. There are some partially successful examples of nation states willing to cede some national sovereignty to international authorities for the global good (for instance, the Montreal Protocol and its extensions to control ozone-depleting substances, the nuclear non-proliferation treaty, or the atmospheric nuclear test-ban treaty). The Kyoto Protocol, even if ratified (currently a dubious prospect), would address only a small fraction of the needed emissions cuts if CO₂ concentrations are to be stabilized below a doubling from pre-industrial levels (much of the primary energy needed in 2050 will have to be mobilized with carbon emissions well below current standards, or huge efforts made to remove the excess¹³) (Fig. 2).

It would require a big increase in 'global-mindedness' on the part of most nations to set up institutions to attempt to control climate and to compensate the losers should the interventions backfire — or even be perceived to have gone awry. Moreover, such an institution would need the resources and authority to make and monitor changes over a century or two — the time it will take the climate system to soak up the bulk of the greenhouse gasses we have injected. Thus, this is the time over which we would continuously need to inject measured amounts of dust in the stratosphere, iron in the oceans^{14,15} or sulphate aerosols into clouds in order to counteract the heat-trapping effects of long-lived constituents such as CO₂.

So the most difficult obstacle in the path of geoengineering may be questionable governance rather than technical uncertainties¹⁶.

Varieties of carbon management

Two broad classes of carbon management can be distinguished. The first includes attempts to manipulate natural biogeochemical processes of carbon removal — so-called 'carbon sinks'¹⁷. The second involves preventing carbon emissions into the atmosphere and instead disposing of it in (one hopes) stable reservoirs. David Keith (see Box 1) suggests that the dividing line between geoengineering and mitigation is when a technology acts by counterbalancing an anthropogenic forcing rather than by reducing it.

Carbon management by manipulating biogeochemical cycles overlaps with geoengineering. Ideas include iron fertilization



Figure 3 Keeping carbon in forests provides a 'double dividend', as primary tropical forests contain good stores of CO₂ and also high biodiversity. But any carbon management scheme must take into account compensation for local people who lose their opportunity to convert the forest to economic product. In addition, monitoring is required to ensure that the carbon stays sequestered and that carbon 'credits' are paid out to the donor to the project over time. (Photo by S. Schneider.)

of the oceans to enhance uptake of carbon by the resulting blooms of phytoplankton, planting vast forests of fast-growing trees to sequester carbon¹⁸ or altering agricultural practices to increase carbon storage in soils¹⁹.

The prevention of carbon emissions that otherwise would have been injected directly into the atmosphere is not geoengineering. Briefly, it includes preservation of primary forests that otherwise might have been cut down (which also helps to preserve biodiversity) (Fig. 3); industrial processing to increase the hydrogen content and remove carbon from fuels such as coal or methane, and subsequent injection of the carbon into storage reservoirs; and using less carbon-intensive energy supply systems and improving energy efficiency. The last two of these, of course, is what has come to be called 'mitigation', and is usually favoured by environmentalists. (The climate policy debate typically argues the costs of mitigation versus adaptation, although geoengineering has been mentioned as a third category from the outset.)

Box 1 Geoengineering

David W. Keith

Geoengineering is planetary-scale environmental engineering, particularly engineering aimed at counteracting the undesired side effects of other human activities¹. The term has usually been applied to proposals for limiting the climatic impact of industrial CO₂ emissions by countervailing measures such as the construction of space-based solar shields. Scale and intent are both central to the common meaning of geoengineering as the following examples demonstrate. First, intent without scale: ornamental gardening is the intentional manipulation of the environment to suit human desires, yet it is not geoengineering because neither the intended nor the realized effect is large-scale. Second, scale without intent: anthropogenic CO₂ emissions will change global climate, yet they are not geoengineering because they are a side effect of the use of fossil fuels to provide energy services.

The distinction between geoengineering and more conventional responses to the CO₂-climate problem is fuzzy. Geoengineering has become a label for technologically overreaching proposals that are omitted from serious consideration in climate assessments. For example, few would object to applying the label to the first pair of examples below, but neither proposal rates serious consideration among climate policy-makers. Conversely, the second pair do receive serious consideration but few would call them geoengineering.

Geoengineering proposals

Enhancing oceanic sinks

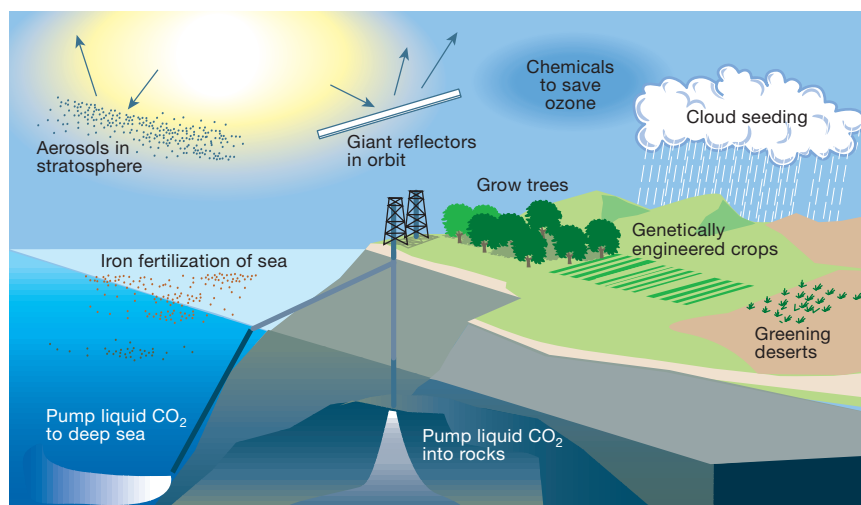
Concept. Fertilizing the 'biological pump' may enhance the flux of carbon into the oceans that maintains the disequilibrium in CO₂ concentration between the atmosphere and the deep ocean. While use of nitrogen and phosphorus has been proposed, iron fertilization is the salient possibility because the ratio of iron addition to carbon fixation is very large (the Fe:C ratio is ~1:10⁴ whereas for N:C it is ~1:6).

Status. Iron-fertilization experiments have produced marked increases in oceanic productivity², and surveys have shown that biological productivity is iron-limited over substantial areas³. Although enhancement of surface productivity is possible, increasing the carbon flux into the deep ocean is highly uncertain — models suggest that even if iron fertilization was used at the largest possible scale the carbon flux would not exceed ~1 GtC yr⁻¹. And problems abound, as iron fertilization could produce anoxia in large regions of the deep ocean.

Shielding some sunlight

Concept. Warming due to anthropogenic greenhouse gases can be countered by deploying systems in the stratosphere or in space that scatter sunlight away from the planet. Stratospheric scatters are much cheaper but entail risks to stratospheric chemistry; space-based systems offer an expensive but clean alteration of the solar 'constant'.

Status. Analysis has shown that it is possible to dramatically reduce the required mass and thus the



Schematic representation of various climate-engineering proposals (courtesy B. Matthews).

cost of both scattering systems⁴. It had long been suggested that changes to the solar constant would compensate only poorly for the climatic effects of increased CO₂, even if mean surface temperature was accurately controlled. But a recent climate model experiment indicates that reduction of solar input can compensate for increased CO₂ with remarkable fidelity⁵.

Ambiguous Cases

Enhancing terrestrial sinks

Concept. Given the substantial human control over the terrestrial biosphere, the large natural carbon fluxes between atmosphere and terrestrial biosphere provide a powerful lever for manipulating atmospheric CO₂. A great diversity of methods have been proposed to exploit this leverage including reforestation and sequestration in agricultural soils via 'zero-till' methods or via the genetic modification of cultivars to enhance lignin content⁶.

Is it geoengineering? Enhancement of terrestrial sinks has been seen as green and low-tech in sharp contrast with geoengineering. The idea has garnered wide support in industry and among environmental organizations. Yet, if implemented at the scale required to capture a significant fraction of emissions, terrestrial sequestration would resemble planetary-scale environmental engineering and may well entail high-tech methods such as genetic modification of crops. The divergent treatment of terrestrial and oceanic sinks illustrates the inconsistencies that pervade discussion of planetary engineering.

Sequestering CO₂

Concept. We may use fossil energy without emissions of CO₂ by first capturing the carbon content of fossil fuels while generating carbon-free energy products such as electricity and hydrogen and then sequestering the resulting CO₂ in geological formations or in the ocean⁷.

Is it geoengineering? The term geoengineering was coined in the 1970s to describe the injection of power-plant CO₂ into the deep ocean. Despite this

etymology it is unclear whether capture and sequestration is rightly classified as geoengineering. It is certainly an end-of-pipe technical fix, but (arguably) injection into geological reservoirs resembles conventional pollution-mitigation technologies more closely than it resembles geoengineering, because it limits emission of CO₂ to the biosphere rather than compensating for emissions after they occur. Put simply: if geological sequestration is end-of-pipe then biological sequestration is beyond-the-pipe.

Commentary

The post-war growth of the earth sciences has been fuelled, in part, by a drive to quantify environmental insults in order to support arguments for their reduction. Yet paradoxically the knowledge gained is increasingly granting us leverage that may be used to deliberately engineer environmental processes at planetary scale. The manipulation of solar flux using stratospheric scatterers is perhaps the best example of this leverage: we could reduce solar input by several per cent — probably sufficient to initiate an ice age — at an annual cost of less than 0.01% of global economic output^{1,4}. As remedies for the CO₂-climate problem, all proposed geoengineering schemes have serious flaws. Nevertheless, I judge it likely that this century will see serious debate about — and perhaps implementation of — deliberate planetary-scale engineering.

David W. Keith is in the Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. (e-mail: Keith@cmu.edu)

1. Keith, D. W. *Annu. Rev. Energy Environ.* **25**, 245–284 (2000).
2. Boyd, P. W. *et al. Nature* **407**, 695–702 (2000).
3. Behrenfeld, M. J. & Kolber, Z. S. *Science* **283**, 840–843 (1999).
4. Teller, E., Wood, L. & Hyde, R. Report no. UCRL-JC-128157 (Lawrence Livermore National Laboratory, Livermore, CA, 1997).
5. Govindasamy, B. & Caldeira, K. *Geophys. Res. Lett.* **27**, 2141–2144 (2000).
6. Rosenberg, N. J., Izaurralde, R. C. & Malone, E. L. *Carbon Sequestration in Soils: Science, Monitoring, and Beyond* (Battelle, Columbus, OH, 1998).
7. Parson, E. A. & Keith, D. W. *Science* **282**, 1053–1054 (1998).

One idea is to build on the chemical industry's existing experience of industrial-scale carbon removal and sequestration. Nitrogen fertilizer, for example, is manufactured when carbon fuels such as natural gas or gasified coal are converted to secondary energy carriers such as hydrogen — although this is not done for the purpose of using the clean-burning hydrogen as a fuel, but rather for chemical processing. And as carbon-intensive fuels such as coal are progressively converted to more hydrogen-based fuels such as methane, carbon dioxide is a by-product that needs to be sequestered in a stable reservoir. The oil industry, too, has long experience with CO₂ sequestration through advanced oil recovery schemes. With one exception, these are not aimed at reducing atmospheric emissions of CO₂. Nonetheless, this experience can be built upon to develop carbon management for climate purposes.

The feasibility of CO₂ sequestration below ground has already been explored at small scales. The Sleipner West offshore platform in the North Sea operated by the Norwegian company Statoil is an interesting experiment in which about 1 million tons of CO₂ annually is stripped out of the natural gas mixture brought out of the earth. The CO₂ is re-injected into an aquifer about 1,000 metres below the ocean surface. As the CO₂ spreads along this geological formation, eventually — perhaps over hundreds of years — it may leak out, but this slow re-injection back into the climate system will avoid the acute build up of CO₂ that would have occurred under normal circumstances. Most interesting, perhaps, is why this first-of-a-kind plant was built: Norway had instituted a tax on carbon emissions of around US\$100 dollar per ton carbon, and it seems that the costs of CO₂ removal and sequestration might be cheaper than paying the tax.

This, of course, is the crux of the climate policy debate: how can we create incentives to put a price on carbon or other heat-trapping gases? Debate rages about whether to provide incentives directly by a carbon tax, or indirectly via targets and timetables (as in the Kyoto Protocol), or via subsidies to companies willing to develop carbon management schemes. But without such incentives, the extent to which technological options will be explored is questionable.

Carbon management thinkers have also suggested that industrial efforts should not be restricted to centralized sources such as power plants or oil platforms, but must consider distributed applications such as transportation systems. Perhaps we will see the development of a few centralized plants to produce hydrogen fuel for zero-emission vehicles. To be cost-effective, such plants would need to be in areas with abundant resources of fossil fuels and adequate storage reservoirs for the waste carbon¹. However, some have questioned whether sequestered carbon will remain buried, and thus whether

carbon credits should be given unless it is proved that the storage is lasting. To eliminate endless debate I propose an inexpensive fix: to add into the injected CO₂ an inert chemical tracer unique to each sequestration site. Thus, the non-detection of this tracer over time would serve to certify that carbon credit is deserved for such sequestration projects.

'Strong' or 'weak' engineering

In 1992 at the Rio Environment Summit the United Nations Framework Convention on Climate Change committed the nations of the world to avoid 'dangerous anthropogenic interference with the climate system'. At that time the thinking was primarily about inadvertent modification. But now it seems that world leaders may have to extend their value judgement about what is 'dangerous' to deliberate interference with the climate system.

Naturally, there is a philosophical debate about whether there is anything ethically wrong with such tinkering. On the one hand, we have progressed from hunting and gathering, it is argued, by increasingly large-scale manipulations of the natural environment. In fact, some environmental writers have despaired that an altered climate is already 'The End of Nature'²⁰. It has been argued that Earth systems engineering and management is the approach "to rationally engineer and manage [the Earth] to provide the requisite functionality", and that this is not a logical transgression of naturalness since the Earth is already "an artefact" of our manipulations²¹. Of course, anything other than a preservation of current structure and function demands a definition of 'improvement', and this judgement will be very different across diverse cultures.

Moreover, there are still areas in the polar regions and deep tropical rainforest where there is "essentially no visible human imprint; where the majority of species have evolved in situ...and where biochemical perturbations are small"²². Such landscapes are not 'artificial' simply because a slight global climatic change has already occurred. We should avoid disrupting them further, rather than using 'light perturbation' as an excuse to turn over the future of all nature to the 'functionality' of the planetary managers.

Given our growing inadvertent impact on the planet, adaptation alone may prove inadequate. But I would prefer to reduce slowly our economic dependence on carbon fuels, rather than to try to counter the potential side effects with centuries of injecting sulphuric acid into the atmosphere or iron into the oceans. Laying stress instead on carbon management, with little manipulation of biogeochemical or energy fluxes in nature, is a much less risky prospect — despite remaining uncertainties about the longevity of deep earth or ocean carbon storage, possible ecological consequences of localized injections of vast quantities of CO₂

in the oceans or the potential damper on global economic development. If preliminary studies prove reasonable, then the cost penalties for closing the industrial cycles by reinserting waste CO₂ back in the earth might be only a few tens of per cent of current energy system costs — something akin to US\$50–100 per ton carbon. The actual costs of various forms of carbon management will be crucial in determining how much climate change we and the unmanaged environment will have to adapt to in the decades ahead. But until national governments cooperate and provide incentives to both producers and users of climate-altering products, the potential for any carbon management enterprise will be restricted and the likelihood of 'dangerous' climatic changes increased.

To me, any form of 'stronger' Earth systems engineering and management is a revision of Rusin and Flit's fantasy of 40 years ago to transform the Earth system to achieve 'improvements in climate'. Those wishing to usurp the province of ancient gods and conjurers should recall the ancient Greeks' warnings about human hubris embodied in the story of Prometheus.

Stephen H. Schneider is in the Department of Biological Sciences and Institute for International Studies, Stanford University, Stanford, California 94305-5020, USA

1. Socolow, R. (ed.) *Fuels Decarbonization and Carbon Sequestration* PU/CEES Rep. No. 302, September 1997 <<http://www.princeton.edu/~ceesdoc/>> (Centre for Engineering and Environmental Studies, Princeton University, 1997).
2. World Commission on Environment and Development. *Our Common Future* (Oxford Univ. Press, Oxford, 1987).
3. Rusin, N. & Flit, L. *Man versus Climate* (Peace Publishers, Moscow, 1960). [Translated from Russian by D. Rottenberg.]
4. Glazovsky, N. F. *The Aral Crisis: The Origin and Possible Way Out* (Naulca, Moscow, 1990).
5. Budyko, M. I. *Climate Changes 244* (American Geophysical Union, Washington DC, 1977). [Translated from Russian.]
6. Kellogg, W. W. & Schneider, S. H. *Science* **186**, 1163–1172 (1974).
7. Marchetti, C. *Clim. Change* **1**, 59–68 (1977).
8. Panel on Policy Implications of Greenhouse Warming, Committee on Science, Engineering, and Public Policy, National Academy of Sciences. *Policy Implications of Greenhouse Warming: Mitigation, Adaptation, and the Science Base* 433–464 (National Academy Press, Washington DC, 1992).
9. Nordhaus, W. D. *Science* **258**, 1315–1319 (1992).
10. Grubb, M., Duong, M. H. & Chapuis, T. in *Integrative Assessment of Mitigation, Impacts and Adaptation to Climate Change* CP-94-9 (eds Nakicenovic, N., Nordhaus, W. D., Richels, R. & Toth, E. L.) 513–534 (International Institute of Applied Systems Analysis, Laxenberg, Austria, 1994).
11. Intergovernmental Panel on Climate Change. *IPCC Working Group I, Second Scientific Assessment, 1995* (Cambridge Univ. Press, Cambridge, 1995).
12. Govindasamy, B. & Caldeira, K. *Geophys. Res. Lett.* **27**, 2141–2144 (2000).
13. Hoffert, M. I. *et al. Nature* **395**, 881–884 (1998).
14. Watson, A. J. *et al. Nature* **371**, 143–145 (1994).
15. Monastersky, R. *Science News* **148**, 220–222 (1995).
16. Schneider, S. H. *Clim. Change* **33**, 291–302 (1996).
17. Intergovernmental Panel on Climate Change. *Land Use, Land-Use Change, and Forestry* (eds Watson, R. T. *et al.*) (Cambridge Univ. Press, Cambridge, 2000).
18. Johansson, T. B., Kelly, H., Reddy, A. K. N. & Williams, R. H. (eds) *Renewable Energy: Sources for Fuels and Electricity* (Island, Washington DC, 1993).
19. Rosenberg, N. J., Izaurralde, R. C. & Malone, E. L. (eds) *Carbon Sequestration in Soils: Science, Monitoring and Beyond* (Proc. St Michaels Workshop, Dec. 1998) (Batelle, Columbus, OH, 1999).
20. McKibben, W. *The End of Nature* (Random House, New York, 1989).
21. Allenby, B. J. *Indust. Ecol.* **2**, 73–93 (1999).
22. Keith, D. W. *IEEE Technol. Soc. Mag.* **19**, 25–28 (2000).

Interfering for the good of a chemical reaction

Stuart A. Rice

Chemistry has, throughout its history, been a science of passive control: having made the conditions as favourable as possible, the chemist is obliged to let the complex reorganization of reactant molecules into products take place unhindered. But what if one could meddle at the very heart of a reaction to affect the dynamics of molecular evolution? Welcome to the world of quantum interference control.

The use of chemistry to support human activity is very old. Think of an archaeological dig, and one immediately thinks of pottery fragments, nails, ancient coins, perhaps metal-tipped weapons or tools — ample evidence that exploitation of chemical operations preceded written history. In the earliest known written records, too, we find references to the preparation of fermented products such as wine and vinegar, and to dyeing and glass-making. It is clear that even in antiquity the amount of practical knowledge concerning chemical processes must have been considerable.

Of course, the relationships between different chemical processes were not yet recognized and organized, and the available knowledge could be transmitted only as practical rules of procedure. But the revolutionary idea that a particular substance can be converted into other substances — by heating, dissolving or mixing — must have spurred the search for new methods to control that transformation and steer it towards the desired products. The science of chemistry developed around that search, and its future will certainly include methods of reaction control unlike anything our forerunners could have imagined.

Over the past two centuries, chemistry has been set on a logical footing; intensive studies of synthetic methodology have given us many ways of generating a vast range of chemical species. The methods that are in current use rely on two fundamental procedures, or a combination of them. The first is the adjustment of a concentration ratio, either of reactants or intermediate species, so as to amplify the yield of the desired product. The second involves adjusting the rates of competing reactions that form different species from the same reactant (or intermediate) so as to enhance the formation of the one that is wanted. In practice, both methods are implemented by altering parameters such as the temperature, pressure, acidity or

solvent composition. What these parameters have in common is that they are measures of the averaged properties of a system with many molecules; their use exemplifies passive control. I use the word 'passive' to convey the notion that the reactant molecules and any surrounding solvent molecules are not subjected to manipulation by external influences during the evolution from reactants to products.

Is such external control possible? We are all familiar with control of the behaviour of machines and other macroscopic systems based on the application of classical mechanics to engineering design. Less familiar are the examples of control of the microscopic behaviour of matter, for which quantum mechanics is the relevant description. However, only in the past few years has it become realistic to think about developing generic tools for active control of product formation in a chemical reaction. By 'active' I mean that, for example, external electromagnetic fields are used to alter the dynamics of molecular evolution during the reaction and thereby generate more or all of a particular product.

Controlling molecular motion

How is it possible to directly control the dynamics of evolution of a molecule? The property of matter that makes this possible is quantum interference, which arises from the fact that the motions of material particles, for example atoms and electrons, can be described in terms of the motion of waves¹. The tool that permits practical exploitation of quantum interference for control of molecular dynamics is the laser. Laser technology now permits the generation of very short pulses of light, shaped pulses of almost any type, pulses with a well-defined phase relationship, very pure monochromatic light fields, and very high intensity light fields.

To understand how quantum interference can help in active control requires one to

abandon some mechanistic ideas associated with the classical description of a chemical reaction. For example, the notion that merely increasing the energy in a particular chemical bond will direct reaction to occur at that bond neglects two characteristic features of molecular behaviour. First, the nature of excitation by light absorption typically spreads the excitation over many bonds and, second, the rate at which energy is redistributed from an excited bond to other parts of the molecule is typically greater than the rate of reaction of that bond. In contrast, control of molecular dynamics through exploitation of quantum interference makes no assumption about the mechanism of energy flow in the molecule.

The quantum description of matter posits that there is a 'de Broglie wave' associated with every particle; it has a wavelength equal to Planck's constant divided by the particle momentum, and like any other wave it has an amplitude and phase. The particle's position cannot be pinpointed as in the classical description; only a probability distribution for the position can be specified. For example, in the lowest energy-state of the hydrogen atom, the electron is distributed spherically about the proton at a mean nonzero distance and with a wavelength equal to the equatorial circumference of the spherical distribution. The probability of finding a particle at any position is proportional to the square of the de Broglie wave amplitude at that position.

As a consequence of the wave character of matter, the possible energy states of electrons in a molecule are not distributed continuously; instead, they can assume only a discrete set of values, just as standing waves in a box can be only of certain frequencies.

These electronic energy states are associated with particular spatial distributions of the electrons of the molecule; electrons occupy a set of molecular orbitals, each of which has a particular spatial distribution and energy. In the lowest (ground) electronic

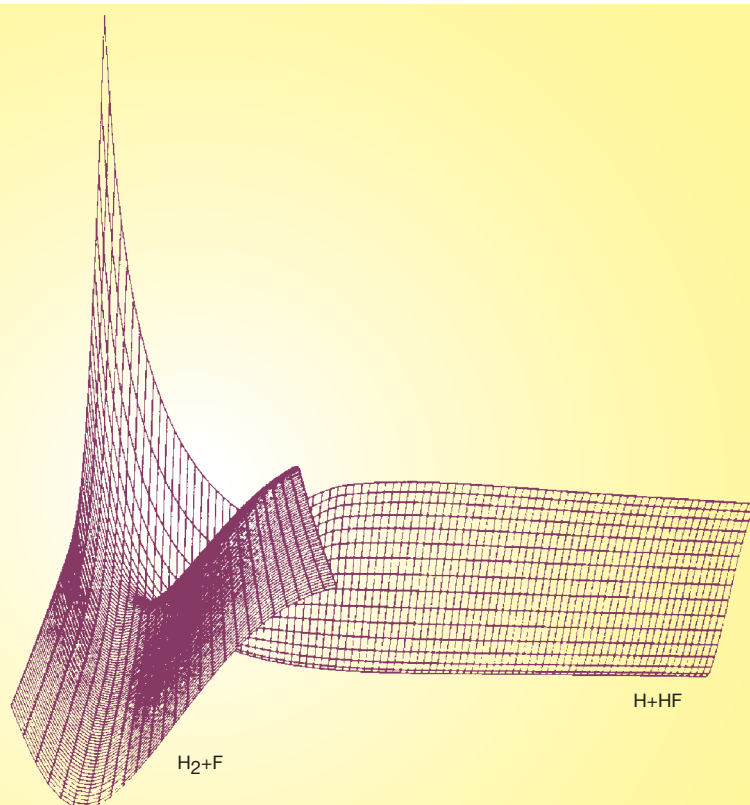


Figure 1 Representation of the potential energy surface for the collinear reaction $F + H_2 \rightarrow H + HF$. The electronic energy of the three interacting atoms is plotted on the vertical axis. The axis pointing out of the page represents the distance between the H_2 molecule and the F atom; the axis pointing to the right represents the distance between the H atom and the HF molecule. The reaction can be envisaged as following a path along which the F atom approaches the H_2 molecule (into the page) until a closest approach arrangement FHH is reached, and then products HF and H separate along the axis pointing to the right. (Adapted from ref. 20.)

state, N electrons fill the lowest energy $N/2$ orbitals, with each orbital having just two electrons.

Excited electronic states of a molecule are created when an electron is transferred from an occupied orbital to an unoccupied orbital, say by the absorption of light. Exciting an electron into a higher energy state corresponds to transferring de Broglie wave amplitude from one state to another. The probability of the transition is proportional to the square of the transition amplitude, that is, the integral of the product of the amplitudes for the two states, the electron displacement and the light field amplitude. It is this relationship between the probability of transition and the amplitude, coupled with the fact that molecular excitation or de-excitation may involve a succession of such transitions, that can generate quantum interference effects, as I shall explain.

There is a strong analogy between interference effects in a molecule and interference of light. Consider, for example, the 'two-slit' experiment familiar to countless schoolchildren, in which there is overlap of coherent waves (waves with a time-independent phase relationship) emanating from

two point sources or slits. The effect of this overlap, viewed on a screen, is to produce brighter regions (wherever the amplitudes of the two waves have the same sign) and darker regions of decreased intensity (where the two waves cancel), thereby generating a spatial pattern of interference fringes. Just this kind of interference pattern can also be observed with electrons diffracted, for instance, by a crystal lattice. More important, there are analogous interference effects in a molecule. Whenever there are two or more independent excitation paths (sequences of transitions) between two states of a molecule, and these independent paths from state A to state B can be excited in a way that yields a constant relationship between their phases — that is, coherently — quantum interference will occur.

The overall probability of a transition from state A to state B is proportional to the square of the sum of the transition amplitudes for all paths that connect these states. Because there is also a phase associated with the amplitude along each excitation path, the differences between the phases of the several paths generate interference that affects the magnitude of the overall transition proba-

bility. Quantum mechanics tells us that it is meaningless to ask which paths are actually taken unless one of them is physically detected in a measurement — at which point the ability to exploit quantum interference between all possible paths is lost. But it is possible to influence the individual paths, and it is the realization of some form of control of the phases of these different excitation paths that permits control of the resulting overall transition probability and hence of the molecular dynamics.

Our visualization of atomic motion in a molecule is aided if we take advantage of the fact that the mass of an electron is about twenty-thousand times smaller than the mass of a typical atomic nucleus. Because of this disparity the electrons move much more rapidly than do the nuclei. The electrons adjust very quickly to any changes in the positions of the nuclei, so the latter in effect move in a field of electrical potential generated by the electrons, the energy varying with the relative positions of the nuclei. This potential field can therefore be mapped with respect to displacements of the nuclei (Fig. 1), and for typical molecules extends into many dimensions according to the number of nuclei; the resulting map is called the potential energy surface for that electronic state. In general, the potential energy surfaces of different electronic states are different.

For many molecular systems the same global potential energy surface describes reactant and product molecules. The reactant molecule is identified with one of the minima of the potential energy surface. Elsewhere on the potential energy surface is another minimum, corresponding to the arrangement of atomic nuclei in the product molecule. A point on the potential energy surface can be used as a classical representation of the state of the molecule. The motion of that point on the potential energy surface can then be used to follow the changes in nuclear positions that accompany the transformation of a reactant into a product molecule. Sometimes the reaction must be followed on a different potential energy surface corresponding to an excited state.

It is not only the electronic motion in a molecule that is important. The small amplitude motions of the nuclei around a particular minimum of the potential energy surface generate a dense but discrete set of vibrational energy levels for that molecule. When fixed in position relative to one another, the overall motion of the nuclei generates a dense but discrete set of rotational energy levels for that molecule. A complete description of the molecule specifies its electronic, vibrational and rotational states.

We can exploit quantum interference to realize control of molecular dynamics in several superficially different ways. For clarity's sake, I will discuss only unimolecular reactions, that is, those situations in which a

single reactant molecule is transformed to several product molecules.

Two-beam interference control

One way is to use two coherent monochromatic laser beams to create simultaneous excitations via different paths between two states of a molecule^{2,3}. This control method is sensitive to the structure of the potential energy surface only through the transition amplitudes between the states of the molecule. These transition amplitudes do depend on the nuclear positions.

Consider the case of a branching chemical reaction in which the excited reactant molecule can form at least two distinct product species. Suppose, for example, that two distinct excitation paths, in which the molecule is excited by absorbing, respectively, one and three photons, can connect the reactant and product states. Imagine that these excitation paths are activated by absorption of photons from two coherent light beams. Because the excitation sources are coherent, so too will be the amplitudes of the excitation transitions — an essential point, if the paths are to interfere. It is possible to modulate the interference between the coherent excitation amplitudes by changing their relative phase, and that is affected by changing the relative phase of their sources. The latter can be achieved by using, for the three- and one-photon excitations, the fundamental and the third harmonic wavelength beams of light derived from a single laser source. When these light beams pass through a gas (Fig. 2a), the dependence of refractive index on wavelength and the distance travelled in the gas can be used to tune the phase difference between them.

Remember the classic two-slit experiment described earlier. In that case the amplitude of light reaching a screen was the sum of the amplitudes of waves from two different paths (the two slits). In the molecular example the excited-state wave amplitude is the sum of the excitation amplitudes generated by two routes that are not distinguished from each other by measurement. The result is, so to speak, an interference pattern in product space, with variations in phase difference giving larger or smaller chances of forming the desired product.

That, at least, is the theory. Execution of the experiments has proven to be very challenging, but was first demonstrated by Robert Gordon and his associates⁴ in studies of the branching reactions $\text{DI} \rightarrow \text{D} + \text{I}$ versus $\text{DI} \rightarrow \text{DI}^+ + \text{electron}$. As shown in Fig. 2b, adjustment of the phase difference between the one- and three-photon excitation paths causes the concentrations of the products to oscillate. The maximum amount of $\text{D} + \text{I}$ is generated (almost exactly) at the same phase as is the minimum amount of DI^+ , hence the ratio of concentrations of the products of the two reactions is controllable by that phase

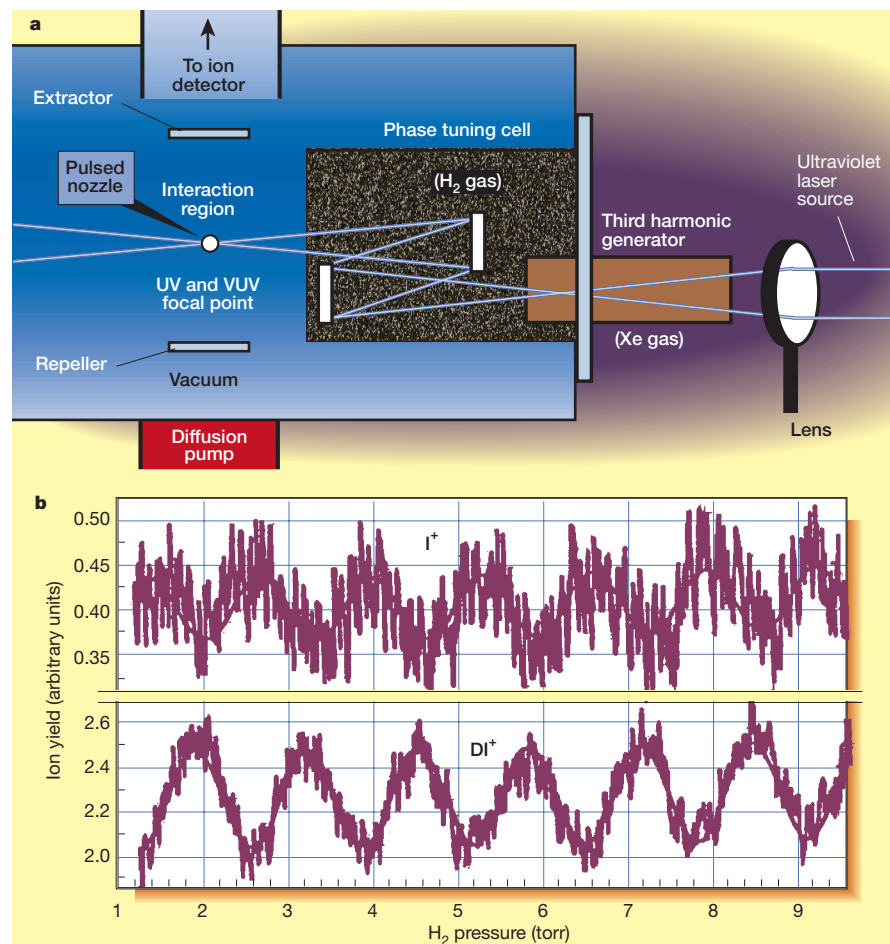


Figure 2 Two-beam interference control. a, Experimental set-up (courtesy of R. J. Gurdon). b, Modulation of the DI^+ and I^+ signals as a function of phase difference between one- and three-photon pathways. The pressure of H_2 is a surrogate for the phase difference. (Adapted from ref. 21.)

difference. A generalization of two-path interference control to multiple-path intense-field control is described in Box 1.

Two-pulse time delay control

A different realization of control of molecular dynamics can be achieved using a variable time delay between two extremely short light pulses⁵. A short light pulse can be represented as a coherent superposition of many monochromatic light waves; the range of the frequencies of the contributing light waves is inversely proportional to the duration of the light pulse. Because many of those frequencies can excite transitions, a very short light pulse applied to a molecule can simultaneously excite many coherent transitions to excited states. The state of the molecule thereby created consists of a coherent superposition of many excited states, corresponding to a superposition of de Broglie waves with many frequencies. A superposition of amplitudes of this type is called a wave-packet state, and its amplitude is usually localized in a small region of the excited-state potential energy surface of the molecule. The motion of the wave packet, and its control, does

depend on the nature of the potential energy surface, and its variation with nuclear geometry. In other words, the way the wave-packet amplitude moves on the potential energy surface is sensitive to the presence of hills and valleys on the surface, respectively representing nuclear arrangements with higher and lower energies. At some time after the excitation pulse, because the frequencies of the components of the wave packet are different, the centre of the wave packet moves to a different position on the potential energy surface and its shape changes.

Consider the case that the excited-state wave packet moves to a new position and is then excited by a second pulse of light to a different potential energy surface. The time difference between the pulses can then be used to place amplitude into selected reaction channels and thereby control the branching to form different reaction products. The fashion by which interference affects this procedure can be seen as follows. Because each of the pulses contains many coherent frequency components, there are very many combinations of frequencies that lead from the initial state to the final state,

each defining a path. The time delay between the pulses controls the relative phases of these combinations, and hence determines when the interference is constructive or destructive. Experimental studies reported by Gustav Gerber and associates⁶ show that in the branching reactions $\text{Na}_2 \rightarrow \text{Na}_2^+ + \text{electron}$ versus $\text{Na}_2 \rightarrow \text{Na}^+ + \text{Na} + \text{electron}$, the ratios of concentrations of the products do indeed depend on the time delay between the pulses (Fig. 3).

Optimizing the control process

Two-beam interference and two-beam time delay are limiting cases of a more general approach to controlling molecular dynamics. Imagine continuously changing the frequency, amplitude and phase distributions throughout the duration of the applied pulse so as to continuously use interference between excitation paths to maximize the yield of a particular reaction product. The shaped pulse that accomplishes this goal is found by starting with the desired product state and the quantum mechanical equations of motion, and then working backwards in time to calculate the time distribution of electric field amplitude, frequency and phase that is required to maximize the product yield. A best (optimal) solution, subject to specific constraints (such as a restriction on the pulse energy or shape) can usually be found^{7,8}. The efficiency with which this pulse enhances the yield of a particular product is determined by the extent of interference between the amplitudes associated with its different spectral and temporal components.

Because the potential energy surfaces of molecules are complicated, the calculated optimal control pulse usually has a complicated distribution of frequencies, amplitudes and phases that change throughout its duration. Fortunately, this pulse optimization method lends itself to experimental automation⁹. Gerber and associates¹⁰, following up the seminal work of Herschel Rabitz and co-workers, have used a learning algorithm to optimize the phase distribution in a light pulse and thereby maximize the relative yield of one or the other of the products of two different bond-breaking reactions of the complicated molecule iron-cyclopentadiene-dicarbonylchloride, $\text{CpFe}(\text{CO})_2\text{Cl}$. This molecule is a metal-ligand complex of the kind that often acts as a catalyst for a biochemical reaction. For the purpose of this demonstration, the important property it has is that the different bonds, iron-cyclopentadiene, iron-carbon monoxide and iron-chlorine, each have different strengths. The experimental results show that the pulsed light incident on a molecule can be optimized to maximize the yield of CpFeCOCl (obtained by cleavage of one of the FeCO bonds of $\text{CpFe}(\text{CO})_2\text{Cl}$) relative to the yield of FeCl (obtained by cleavage of both FeCO bonds and the FeCp bond), or to minimize that yield.

Limits and applications

Active control of molecular dynamics for the purpose of optimizing product selectivity represents a paradigm shift in thinking about chemical synthesis. At first the suggestion that an optical field could control the evolution of a molecule by actively influencing the temporal development of the molecular dynamics, and thereby the choice of products in a chemical reaction, was greatly resisted. This resistance arose, in part, from the entrenched practice of describing reaction pathways using concepts from classical mechanics. It was augmented by the strongly held view that intramolecular energy redistribution is always so much faster than unimolecular fragmentation or isomerization that the ratio of products formed in a branching reaction is determined entirely by the statistical properties of the energy distribution in the reactant molecule. Furthermore, many researchers believed that generation of optical fields with the spectral and temporal characteristics required was not possible. It is the combined progress in introducing subtle ways of exploiting quantum interference, the explosive advances in laser technology and

the successful execution of very complex experiments that has led to the current status of the field.

Despite its successful demonstration, it is unlikely that active control methods can replace any significant fraction of the well-honed synthetic procedures now in use. For one thing, most chemical reactions involve collisions between two molecules. Active control of the product yield in a bimolecular reaction requires control of the dynamics of the three-dimensional collisions of polyatomic reactants. Methods for achieving that control are in the earliest stage of development¹¹. In addition, most chemical syntheses are carried out in solution, and there remain theoretical and experimental uncertainties about the applicability of the currently available active control methods to reactions in solution. Perhaps most important, the cost of making precisely shaped laser pulses is high relative to the costs associated with current technology for carrying out reactions in solution.

Looking ahead

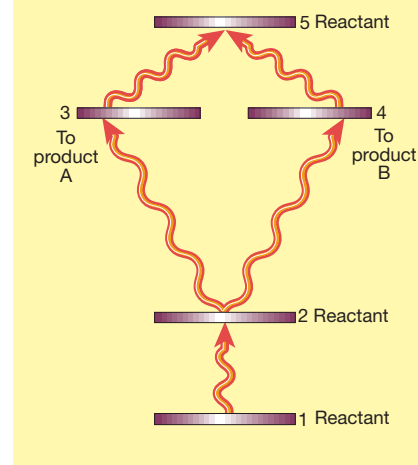
To what extent can we expect quantum interference control of molecular dynamics

Box 1 Multiple-path intense-field interference control

One of the most promising and most efficient methods for controlling molecular dynamics is based on the observation that interference between multiple paths connecting the same initial and final states applies to all kinds of transitions^{17–19}. Suppose, for example, that the relevant reactant and product states are as shown in the figure opposite. States 3 and 4 represent different products generated from the reactant, and these states have the same energy. One of the tasks of the control method is to selectively generate the product corresponding to state 3 rather than that corresponding to state 4, or vice versa. Imagine exciting the reactant molecule with a pulse of light from state 1 to state 2, and then with another pulse of light from state 2 to the product states 3 and 4. The product branching ratio will be determined by the ratio of the intensities of the transitions $2 \rightarrow 3$ and $2 \rightarrow 4$.

But now suppose the molecule is also irradiated with a third long-duration, strong pulse that has the correct wavelength to cause transitions between states 3 and 4 and a further reactant state 5. This pulse creates two coherent superpositions of states: 3 and 5, and 4 and 5. So now we have an indirect pathway to reaction, one that takes advantage of the cycling of amplitude in the coherent superposition of states 3 and 5 and of that in the coherent superposition of states 4 and 5. The frequency of that cycling increases as the pulse strength increases. The amplitudes associated with each of the multiply cycled paths, for example $1 \rightarrow 2 \rightarrow 3 \rightarrow 5 \rightarrow 3 \rightarrow 5 \rightarrow \dots \rightarrow 3$ and $1 \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow 4 \rightarrow 5 \rightarrow \dots \rightarrow 4$, can interfere with the amplitudes associated with the direct paths to products 3 and 4 respectively.

The pulse sequence required is counterintuitive. The pulsed field promoting the transitions $3 \rightarrow 5$ and $4 \rightarrow 5$ must be on while the pulsed fields that promote the transitions $1 \rightarrow 2$, $2 \rightarrow 3$ and $1 \rightarrow 2$, $2 \rightarrow 4$ are applied. Also, the pulsed field that promotes the transitions $2 \rightarrow 3$ and $2 \rightarrow 4$ must be applied before (but overlap with) the pulsed field that promotes the transition $1 \rightarrow 2$ so that when the coherent superposition states are subsequently coupled to state 1, new states are formed from which the pulse cannot transfer population to the intermediate state 2. Rather, the population is channelled directly into the product states. The relative populations of the product states are inversely related to the transition strengths between states 3 and 5 and states 4 and 5, so that by making different choices for state 5 one can control the ratio between the products.



to find practical application?

I believe that laser technology will continue to advance and provide the tools necessary to implement theoretically sound active control methods. To a large extent then, the future for applications of active control of product will depend on the resolution of several fundamental theoretical issues. For instance, is there a limit to the attainable control of quantum dynamical processes? How does the efficiency of a control process depend on the number of degrees of freedom of the molecule? Is it possible to control bimolecular reactions? To what extent is the control of molecular dynamics possible when dissipation must be accounted for, as for a reaction in the liquid phase? How sensitive is the control field to fluctuations in the experimental environment and (from the point of view of prediction) to uncertainties in our knowledge of the molecular potential energy surface?

Partial answers to each of these questions are available, but better answers are needed. For example, the most powerful mathematical theorems available¹ concerning the limit to attainable control of molecular dynamics apply to the case when the possible states of the system are discrete and have energies that are all different. In that case, if the path from the initial to the final state is unrestricted, meaning that any sequence of states can be used to connect the initial and final states, it is possible in principle to transfer 100% of the molecules to any selected final state. Because most chemical reactions involve molecules that have both states with energies separated from one another by discrete amounts and states distributed continuously with respect to their energy, these theorems provide little guidance as to the limit of controllability of product selection.

Nor does demonstration that a light pulse can be tailored to optimize the yield of one product in a branching reaction establish the limit to the efficiency obtainable. Introducing adaptive learning in the active control process is one means of compensating for uncertainties in our knowledge.

The size of a molecule is relevant to the design of the process that controls its molecular dynamics for the following reason. As the number of atoms in a molecule increases so does both the number of energy levels per unit of energy and the complexity of the atomic motions associated with those energy levels. For active control of the dynamics of a polyatomic molecule to succeed, the coherence of the subset of excited states prepared by the laser pulse must not be rapidly destroyed by interactions with other levels of the molecule. The most recent experimental evidence (for which there is not yet a satisfactory theoretical description) strongly suggests that in many polyatomic molecules this loss of coherence can be relatively slow

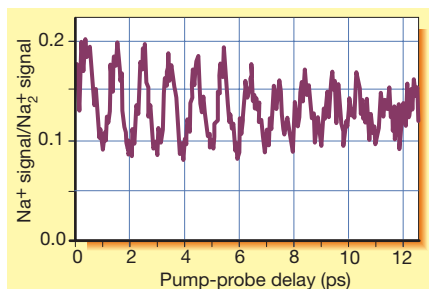


Figure 3 Ratio of Na^+ and Na_2^+ ion signals as a function of pump-probe delay time. (Adapted from ref. 22.)

on the timescale of the control laser pulse¹², increasing the chances of success.

What about active control of the product yield in a bimolecular reaction? Control of the dynamics of a bimolecular collision can in principle be achieved by arranging for interference between different pathways connecting the same initial and final states of the colliding molecules. This task is very demanding, but there are indications from theory¹¹ that some special choices of the centre-of-mass motions of the reactants will generate the needed interference.

Reactant molecules in solution are constantly battered by solvent molecules, and these collisions ultimately destroy the coherence of the excited states necessary to actively control product formation. The key to active control of reactions in solution is reducing the influence of collisions between reactant and solvent molecules on the timescale needed for establishment of the interference that generates selectivity of product formation. The little available information concerning the timescale on which reactant-solvent collisions destroy excitation coherence suggests that in many cases it is several picoseconds. Because the lengths of optimized control pulses are around half a picosecond, there is a window for successful competition between retaining the coherence necessary for active control and destruction of that coherence.

Because the underlying principles of quantum interference control of molecular dynamics are broadly applicable, it is likely that applications in a variety of other fields will be developed in the near future, probably before practical applications in chemistry. One field in which this is already starting to happen is optoelectronics. For example, it is possible to use laser light to control the state of a single quantum dot (a nanometre-sized sample of metal or semiconductor with widely spaced energy levels), and interference effects can thereby be used to generate a fast semiconductor optical switch^{13–15}. It has also been shown that shaped light pulses can greatly enhance the yield of soft X-rays from gaseous argon

and to alter in a programmed way the frequency distribution of the X-rays¹⁶.

Commercial prospects

Although the cost of producing laser light is very high, that may not rule out the use of active control methods for commercial production of chemicals. At present, photochemistry initiated by incoherent light is used industrially in only a few cases for the production of high-value intermediate and end-product chemicals, usually pharmaceuticals or perfumes. Because the cost of producing laser light is dropping, the possible advantages that accrue from the cleanliness of enhanced product generation by controlled laser pulses, for example by reduction in the need for expensive separation processes, may yet change the economic balance for some commercial products.

Going further into the realms of speculation, one can imagine developing active control methods for generating selective reactions of high-value added molecules, and for the design of new (not naturally occurring) biologically active molecules.

Peering still further into the future, if we can achieve active control of the selective reaction of biologically active molecules in solution, one can imagine developing a medical procedure that, using a fibre optic probe *in situ*, destroys by controlled reaction a specific molecule essential for the progression of a disease. Pipe dreams, perhaps; but when one sees how far we have come since the mix-it-heat-it-and-see days of early chemistry, maybe not so impossible a project after all.

Stuart A. Rice is at the James Franck Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA
(e-mail: s-rice@uchicago.edu)

1. Rice, S. A. & Zhao, M. *Optical Control of Molecular Dynamics* (Wiley, New York, 2000).
2. Shapiro, M. & Brumer, P. *J. Chem. Phys.* **84**, 4103–4104 (1986).
3. Shapiro, M. & Brumer, P. *J. Chem. Phys.* **97**, 6259–6261 (1992).
4. Zhu, L. *et al. Science* **270**, 77–80 (1995).
5. Tannor, D. & Rice, S. A. *J. Chem. Phys.* **83**, 5013–5018 (1985).
6. Baumert, T., Grossert, M., Thalweiser R. & Gerber, G. *Phys. Rev. Lett.* **67**, 3753–3756 (1991).
7. Peirce, A. P., Daleh, M. A. & Rabitz, H. *Phys. Rev. A* **37**, 4950–4964 (1988).
8. Kosloff, R., Rice, S. A., Gaspard, P., Tersigni, S. & Tannor, D. *Chem. Phys.* **139**, 201–220 (1989).
9. Judson, R. S. & Rabitz, H. *Phys. Rev. Lett.* **68**, 1500–1503 (1992).
10. Assion, A. *et al. Science* **282**, 919–922 (1998).
11. Abrashkevich, A., Shapiro, M. & Brumer, P. *Phys. Rev. Lett.* **81**, 3789–3792 (1998).
12. Grubele, M. *Adv. Chem. Phys.* **114**, 193–261 (2000).
13. Bodedo, N. H. *et al. Science* **282**, 1743 (1998).
14. Kuriziki, M., Shapiro, M. & Brumer, P. *Phys. Rev. B* **39**, 3435–3437 (1989).
15. Dupont, E., Corkum, P. B., Liu, H. C., Buchanan, M. & Wasilewski, Z. R. *Phys. Rev. Lett.* **74**, 3596–3599 (1995).
16. Bartels, R. *et al. Nature* **406**, 164–166 (2000).
17. Shapiro, M., Chen, Z. D. & Brumer, P. *Chem. Phys.* **217**, 325–340 (1997).
18. Kobrak, M. N. & Rice, S. A. *Phys. Rev. A* **57**, 2885–2894 (1998).
19. Kobrak, M. N. & Rice, S. A. *J. Chem. Phys.* **109**, 1–10 (1998).
20. Schaefer, H. F. *The Electronic Structure of Atoms and Molecules* (Addison-Wesley, 1972).
21. Zhu, L. *et al. Phys. Rev. Lett.* **79**, 4108–4111 (1997).
22. Baumert, T., Helbing, J. & Gerber, G. *Adv. Chem. Phys.* **101**, 47–82 (1997).

Future optical and infrared telescopes

Roger Angel

New telescopes, new detectors and new regions of the electromagnetic spectrum have often revealed totally unsuspected aspects of the Universe. How can we maximize the chances of such serendipity in the future?

Much of my time, like that of other researchers, is spent writing proposals to funding agencies. These generally start out by stating the scientific goals and expected results, and even if we think and hope that unforeseen discoveries are likely, they are not a good selling point. It is refreshing, then, to think specifically about paths to maximize the unexpected. By definition, we cannot foresee the unforeseeable, but we can hope to recognize the paths which in the past led to such discoveries, and to make sure we leave them open in the future.

Let us take as examples unforeseen discoveries that made for dramatic advances in cosmology in the past century. Consider the careful spectroscopic observations of nebulae started around 1912 by Vesto Slipher, using the modest 24-inch (61-cm) telescope of Lowell observatory. The recession velocities of thousands of kilometres per second that he found were completely unexpected, and only in 1917 did deSitter begin to understand them as galaxies in an expanding Universe. Dark matter, a still unidentified form that outweighs all normal matter in the Universe and controls the condensation of matter after the Big Bang, was discovered by Fritz Zwicky and Vera Rubin during spectroscopic observations to weigh galaxies. The rapid motions reflected a gravitational pull too large to be accounted for by ordinary matter in stars and interstellar material. The microwave background radiation, the cooled light of the Big Bang itself, was completely unforeseen by its Nobel prize-winning discoverers, Arno Penzias and Robert Wilson (although not by others).

What common threads do we find in these and the many other unexpected discoveries in astronomy? Very often the observers were pushing new equipment to its limits to accomplish a highly focused goal unrelated to the actual discovery. Slipher was hoping that the diffuse nebulae might turn out to be nearby planetary systems in formation. Zwicky and Rubin were using the largest telescopes to see fainter objects than ever before. Penzias and Wilson were testing a highly sensitive new type of radio telescope for communications.

More generally, unforeseen discoveries are made when the range of observations is enlarged. This happens whenever new parts of the electromagnetic spectrum are opened for observation. New vistas are also opened when larger telescopes or better detectors allow us to reach fainter objects or bigger samples, or to study brighter ones in more detail or with better time resolution. A good discussion of these topics is given by Martin Harwit in his book *Cosmic Discovery*¹.

We cannot discuss future telescopes before pointing out a sea change in astronomy that is likely to drive new instruments in the coming decades. This is the exploration of exoplanets, the planets of other stars. The existence of other living worlds like our own in the Universe has been the subject of speculation for centuries. But telescopes have not been powerful enough to discern extra-solar planets. In the Solar System, the Earth already occupies the prime location and our neighbours appear too hot or too cold. Now at last with clear evidence that giant planets exist around other stars, we have the technical possibility and the incentive to build radical, new telescopes to study them. With them we should be able to find planets even as small as Earth, and to search their spectra for the biochemical signs of life. The potential for unforeseen discovery is enormous. Already the first momentous discoveries have shown completely unpredicted phenomena: planets of Earth's mass orbiting a pulsar, and planets with the mass of Jupiter orbiting closer to their stars than does Mercury to the Sun, a place far too hot for them to form. Who knows what surprises are in store as we begin to explore new worlds and to turn on the rest of the Universe the new telescopes with the sensitivity to see exoplanets?

Looking ahead at the new technology that should lead to unforeseen discovery, we see first a wave of key advances coming in ground-based telescopes in the next decade, the fruits of investments in the past decade. Space astronomy's next big leap will come later, the result of the current period of design and development of concepts for big, cold telescopes to study exoplanets and the

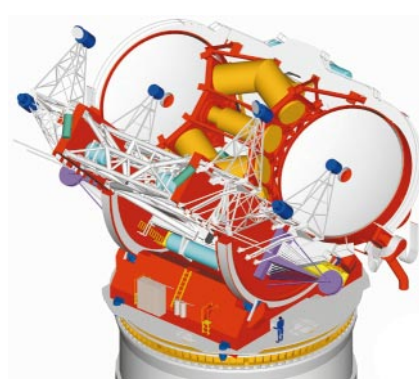


Figure 1 The Large Binocular Telescope, with two 8.4-m mirrors side by side. (Drawing by European Industrial Engineering, Mestre, Italy.)

formation of galaxies. Still further in the future, we may see huge space telescopes of gossamer construction, if the ideas just being formulated finally are realized. I will deal in turn with these areas.

Ground telescopes come of age

During the 40 years after the Palomar 200-inch (5-metre) telescope came into operation the power of optical telescopes was increased enormously, even though none significantly bigger was built. The gains were made by improving detector sensitivity 100-fold to reach the fundamental limit set by photon noise, by extending detector sensitivity into the infrared and by multiplexing so that dozens or hundreds of objects could be analysed at once. But as these advances have reached their limits, the past decade has seen the construction of a new generation of much larger telescopes. Starting with the two 10-m Keck telescopes, there are now around a dozen of size 6.5–10 m in operation or under construction. With such an increase in collecting area, some profound unforeseen discoveries can be expected. This will be especially true when these telescopes are able to remove atmospheric blurring and are linked together as interferometers. Up until now, astronomers have had to choose between small-aperture telescopes in space

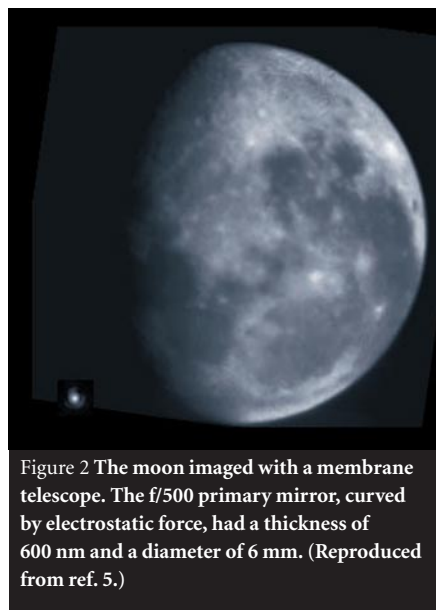


Figure 2 The moon imaged with a membrane telescope. The f/500 primary mirror, curved by electrostatic force, had a thickness of 600 nm and a diameter of 6 mm. (Reproduced from ref. 5.)

that are free from blurring (the Hubble Space Telescope or HST) or large-aperture telescopes on the ground with blurring. Adaptive optics is a new technique that will give both advantages at once². In fact, because the natural limit to resolution set by diffraction improves in proportion to aperture, the bigger ground telescopes will be several times sharper than HST.

The key element in adaptive optics is a mirror whose shape can be altered rapidly in response to the measured atmospheric distortion. By giving the mirror equal but opposite distortion, the original image sharpness is restored. It has been difficult to make the measurement and correction fast enough to keep up with the constantly changing turbulence, but today's detectors and computers make it possible. Already correction of bright objects has been accomplished, and in a few years we should start to see sharp images of even the faintest objects corrected at infrared wavelengths. When a target itself is too faint to allow fast measurement of atmospheric distortion, an artificial star created by a laser searchlight may be used as a surrogate. Experimental laser systems are in operation, although the combination of exquisite tuning and high power needed to excite scattering very high in the atmosphere is proving elusive.

Once adaptive optics are in place another technique to increase the scope for discovery becomes possible. This is interferometry, long used by radio astronomers, which relies on combining the waves from separate telescopes. By measuring the strengthening and weakening of intensity as the crests and troughs reinforce or cancel out, images with greatly increased angular resolution are obtained. Interferometry can be extended to optical wavelengths for very high resolution imaging, but sensitivity is poor unless all the waves from each dish add together coherent-

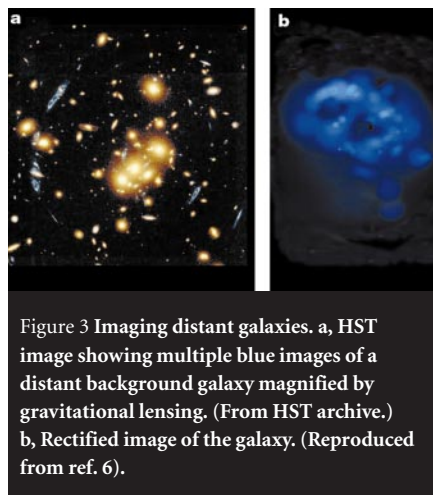


Figure 3 Imaging distant galaxies. a, HST image showing multiple blue images of a distant background galaxy magnified by gravitational lensing. (From HST archive.) b, Rectified image of the galaxy. (Reproduced from ref. 6).

ly. In the presence of atmospheric distortion, coherent addition is possible only when very small apertures are used. But once the full apertures of the large telescopes are corrected with adaptive optics and used as interferometer elements, very faint objects will become accessible.

All of the largest telescopes have been built as multiple units in anticipation of interferometry: Europe's Very Large Telescope (VLT) consists of four telescopes while the Keck and Large Binocular Telescope (LBT) each have two (Fig. 1). Because the LBT mirrors are on a common mounting and closely spaced, as an interferometer it will act like a 22-m telescope with a resolution of 10 milliarcseconds and high infrared sensitivity to complex, faint sources over a wide field of view. It will be able to find and analyse the thermal spectrum emitted by self-luminous giant planets, younger versions of our own Jupiter. The VLT and Keck telescopes are separately mounted and widely spaced (up to 100 m) for resolution as high as 2 milliarcseconds over a field of ~1 arcsecond, but at the price of lower sensitivity and some ambiguity in image structure. They should be able to resolve those Jupiter-sized exoplanets mentioned above that are so close to their stars that they are red hot.

Now that the construction phase of the current generation of big telescopes is winding down, their designers are looking at ways to build bigger optical telescopes. Californians are studying a single telescope of 30-m diameter (the California Extremely Large Telescope); Europeans one of 100 m (the Overwhelmingly Large telescope). Arizona's focus is on an imaging interferometer with two 20-m moving telescopes, the 20/20 telescope. With adaptive optics and interferometry, their sensitivity and resolution can potentially far outstrip what we have now. The tasks we can foresee for such telescopes include observations of the very early Universe at much higher resolution and sensitivity than possible with HST, and the detailed spectroscopic study of Jupiter-like

exoplanets. If the closest stars have Earth-like planets, they should be visible in long exposures with the 20/20 telescope. In addition, the prospect of unforeseen discovery is a powerful motivator.

New discoveries are likely not only from such larger telescopes and higher-resolution images, but also from automated analysis of deep-sky images. In the past, photographic plates from 1.2-m telescopes covering 6° of the sky on a side have provided a rich source of data. Now electronic detectors are greatly increasing accuracy, sensitivity and spectral range. Two digital all-sky surveys with 1.2- and 2.5-m telescopes are in progress now, with charge-coupled device detectors for optical wavelengths in a mosaic of area ~0.1 m² and near-infrared detectors of 0.001 m². The coming decade should see the construction of the 8-m Large Synoptic Survey Telescope with 3° field (0.5-m² detector illuminated at f/1.25)³. The data rate will be prodigious, with the night sky repeatedly mapped every four clear nights, creating 10 terabytes of data per map. Two key goals will be finding and measuring the orbits of asteroids that could hit the Earth as small as 300 m, and mapping the spatial distribution of dark matter on large scales. Dark matter is detected through the small but systematic distortion it induces in the shapes of background galaxies. The data reduction for these tasks is big, but well defined. The challenge for computer gurus will be to spot the unforeseen — for example, new types of erratic variable objects or subtle large-scale correlations in data.

The limitless potential of space

Despite the evolution of much more powerful ground telescopes, space still holds unique and fundamental advantages. Complete freedom from atmospheric distortion in space will remain a major asset. Even when laser guide stars are used for adaptive correction on the ground, a nearby background star is still needed to sense overall jitter of the sharpened image. Stars are dense enough for infrared imaging over most of the sky, but optical imaging, which demands more accurate and faster stabilization because of the shorter wavelengths, will be restricted to fields near a bright star. Tomographic, three-dimensional mapping of atmospheric turbulence with multiple laser guide stars may be able to relieve this optical restriction and open wide-field correction. Although possible in principle, this will be very challenging.

The second unique capability of telescopes in space is for observations in the blocked ultraviolet and infrared spectral regions. Even where it transmits in the infrared, the atmosphere is bright from heat and molecular emission. Additional background heat comes from the ground telescope optics themselves, which cannot be cooled because that would cause them to frost. Space telescopes can be



Figure 4 Evolution of infrared space telescopes. Shown to relative scale are SIRTF, a 0.85-m cryogenic telescope to be launched in 2002, and a concept for the 8-m NGST. Both will orbit far from the Earth, with thermal shields on the sunward side.

made extremely cold for a million times reduction in sky background.

The current generation of space telescopes has already shown how these advantages can be exploited. HST records wide-field optical images typically ten times sharper than for uncorrected ground-based telescopes, and reaches far into the vacuum ultraviolet. Its capability has been steadily improved by astronauts. For example, a new instrument was added to extend imaging and spectroscopic capability to the near infrared, where the natural sky background from sunlit interplanetary dust is much darker than on Earth. HST is likely to remain the only large optical telescope in space for astronomy for at least the next decade. Its dominance will be weakened by a strong challenge from ground telescopes, but perhaps that will allow time for more speculative use that could lead to unforeseen discovery.

HST's mirrors are too warm to be useful further into the infrared where on the ground the atmosphere completely blocks transmission. Here three other space telescopes, the Infrared Astronomical Satellite, Cosmic Background Explorer and Infrared Space Observatory, all cooled with liquid helium to within a few degrees of absolute zero, have revealed the very cold universe. New discoveries can be expected from the Space Infrared Telescope Facility (SIRTF), a new helium-cooled telescope with 0.85-m aperture and more advanced detectors, which is scheduled for launch in 2002. It will reach to 160- μm wavelength.

Beyond these telescopes, the potential for future space astronomy is limitless. In principle, arbitrarily high sensitivity and resolution can be achieved at any desired wavelength, by making the telescope large, accurate and cold enough. Space is an ideal environment, because not only is atmospheric degradation removed, but so also are gravity and wind. This is a huge advantage because, under such forces, big objects bend more than small ones, whereas no matter

how big a telescope is made, the mirror surface accuracy cannot be relaxed if the waves are to be reflected to a sharp focus.

What might a future telescope look like, built for operation in orbits far from the Earth where distorting forces are minimal? The actual working part of a telescope mirror is the top layer of the reflecting surface, which is only a few hundred aluminium or silver atoms deep. Thus there is the possibility in space of using extremely thin membrane mirrors covering hundreds or thousands of square metres, if a way can be found to hold the shape. One way is to stretch a membrane flat from the edge, and then to curve it very slightly by electrostatic force. An experimental telescope made in this way from a membrane 600-nm thick (a few thousand atoms) was used to record the image shown in Fig. 2. While very small, this mirror paves the way for larger-membrane telescopes, with tests planned first on the ground and then in space.

For such gossamer telescopes in space the main disturbance comes from sunlight, which both warms spacecraft and pushes them. Light pressure was strong enough to disturb noticeably the orbit of the 100-foot (30.5-m) Echo balloon satellite launched in 1960 (only 4 years after Sputnik), and would blow a 600-nm membrane right out of the Solar System. Gossamer telescopes thus would have to be shielded from sunlight, but the shields must be lightweight too, and held off the telescope somehow, for solar pressure would push a freely orbiting shield into collision with the shielded mirror. Future gossamer telescopes may look more like sailing ships than the telescopes we now know.

Suppose at some time we find a planet that shows clear chemical evidence of life. Could we image it? The largest space telescopes we can conceive do not use mirrors at all and are extremely heavy, focusing light by gravitational bending. Figure 3a shows how multiple images of the same very distant galaxy are formed and distorted into magnified arcs by an intervening blob of dark matter, traced out by the brighter galaxies held in its gravitational field. The distortion was removed to yield the image shown in Fig. 3b. Our deepest views of the distant Universe will probably rely on finding chance alignments like this to aid our biggest ground and space telescopes. There is no chance that a nearby black hole will lie conveniently in front of and magnify an interesting planet, but we could use the Sun as a gravitational lens by sending a detector out of the Solar System in the opposite direction to the planet. Once it is far enough away, more than 600 times the radius of the Earth's orbit, light from the planet bent around the Sun will be brought to a focus, forming a large image. Taking a detector to such distance may sound difficult, but it is still far easier than interstellar travel to even the nearest star, 1,000-times further away.

The next space telescopes

Limited only by the laws of physics and our imagination, we can see this enormous potential for space telescopes. But what path do we now take? We must be guided by a judicious mix of scientific goals and technological opportunity. The astronomy community has identified two key scientific goals to shape the development of more powerful space



Figure 5 A 2-m hexagonal concave glass mirror is lifted clear after separation from the support on which it was thinned and polished. It is 2 mm thick and weighs just 30 lb (13.6 kg)⁷. (Image courtesy of J. Burge.)

telescopes. One is to probe deeply the highly redshifted, early Universe of which we have caught the first glimpse with HST. It is best observed in the virtually unexplored 2–5- μm region of the electromagnetic spectrum where the natural sky background is darkest. The other goal is the exploration of other planetary systems, with particular emphasis on looking for Earth-like planets, which will be restricted to only the very nearest stars from the ground. This is best attempted from space by detection of the infrared heat emitted by planets (10–20 μm) rather than the starlight they reflect. Their thermal emission is much brighter relative to the star, and the spectroscopic signatures of key diagnostic molecules, water and ozone, could be very strong, as they are in Earth's spectrum. Ozone would be of enormous interest, as in our atmosphere it is of biological origin⁴.

Remarkably enough, both these major scientific objectives involve similar technical advances in the previously neglected shorter wavelengths of the thermal infrared spectrum. Both need mirror temperatures in the range of 30–100 K, which should be reachable without expendable refrigerants by carefully shielding against solar heating and by lofting so far from the warm Earth that it appears no bigger than the Moon. Mirrors even larger than HST's 2.4 m are desirable in both cases, once the technology can be developed to make them. The extreme lightweighting of membrane structure is not yet needed, but the passive, rigid mirrors that are the current standard for all space telescopes will no longer be adequate. We will have to learn how to transfer to space the key technology that has made the new, big ground telescopes possible — namely, active control to correct distortion caused by mechanical or thermal disturbance. Ground telescopes automatically check star images as fast as every minute and make appropriate corrections of alignment and mirror distortion to maintain optical quality.

NASA administrator Dan Goldin has encouraged astronomers and aerospace companies to set ambitious goals. Detailed studies are being made of a telescope envisaged to succeed HST, the Next Generation Space Telescope (NGST), with a primary mirror 8 m in diameter operating at a temperature of 50 K (Fig. 4). Because the practical limit for a telescope launched in one piece is 4–6 m, depending on the size of future rocket fairings, it would be necessary to build the mirror and telescope in space. Planning has so far focused on automatic deployment from an unmanned payload. The whole spacecraft will have to weigh a lot less than HST, even though the telescope has ten times the mirror area, if it is to be carried a million miles from the Earth.

The mind-stretching NGST goal has been a most effective way to focus research in universities, industry and government centres. New technologies have been identified, and development work is going on in many areas.

As an example, Jim Burge at the University of Arizona has been devising ways to take the mature technology used at the Mirror Lab for 8-m glass mirrors for the ground (weighing 18 tons) and produce an 8-m mirror in space weighing less than 1 ton. The idea, demonstrated 4 years ago at the 0.5-m scale, is to keep only the first couple of millimetres of the reflecting glass surface, and replace the rest with an ultra-lightweight carbon-fibre truss. Figure 5 shows the 2-m diameter, 2-mm thick facesheet that will be connected to a truss by 166 active screw actuators. The whole prototype mirror assembly is one-twentieth the weight of the 2.4-m HST mirror.

Although ultra-lightweight, active mirrors are vital, they are but one of a range of challenging new technologies being studied for NGST. For example, passive cooling to cryogenic temperature is new, and in order to cool an 8-m telescope to 50 K the sunshield shown in Fig. 4 must be the size of a tennis court. It must be such a good insulator that less than one-thousandth of the heat from the full blast of the Sun is allowed to leak through to be radiated into space. If as currently conceived, the huge, flimsy NGST structure and sunshade is to deploy automatically, how do you test beforehand in the Earth's gravity to be sure it will work?

An even more ambitious target has been set for the investigation of Earth-like planets. The basic approach is to detect their heat with a special form of interferometer in space. The beams from the different mirrors must be brought together with an accuracy of a nanometre despite tens of metres separation. Such accuracy is needed to fully cancel out the star's electromagnetic waves by maintaining the crests and troughs exactly out of phase, while leaving the planet image intact, a technique known as nulling interferometry. A mission called Terrestrial Planet Finder (TPF) is envisaged to explore not just the nearest few stars, but over a hundred sun-like stars out to a distance of 50 light years. This would require four or more cryogenic 4-m telescopes spread out over 100 m. They could be connected mechanically or orbited separately in formation.

Reliable and affordable solutions

Now that these scientific goals and the range of technical challenges required to accomplish them are identified, we must find the best path to proving solutions that they will work reliably, and will be affordable. We want our telescopes to make unforeseen discoveries, but we do not want to discover unforeseen flaws! I believe the best way is an evolutionary process based on intermediate missions, each aimed at three goals. First, testing out one or more of the key technologies needed for the full-scale NGST and TPF missions; second, proving that new technology can be delivered affordably; and third, using it for astronomical exploration.

The last requirement I see as crucial. The only way to find unanticipated flaws is by attempting challenging observations in space. Because each of the multiple technological developments for NGST and TPF is profound, any one of them alone could be used to open new scientific frontiers. Consider what might be done relatively soon. Even an old-fashioned, rigid telescope, if taken far from the Earth and passively cooled to 30 K, could open a new window to the high-redshift Universe. A warm, 4-m, ultra-lightweight telescope could test autonomous active control for high-contrast images and see further than HST. A short nulling interferometer with proven, 1-m class cryogenic mirrors in a distant cold orbit could already find and obtain images and spectra of any Earth-like planets around the very nearest stars.

Such missions could also be configured to test solutions to what may be the most difficult problem, namely how to endow remote, cold observatories with HST-like longevity and productivity. If there were no possibility of repair and refurbishment, any big space telescope of HST-level or greater complexity would inevitably be risky and become dated. Space Station provides a natural solution, as a base where complete large telescopes and spacecraft could be assembled and tested in the gravity-free space environment. With solar cells deployed, the structure would be pretty flimsy, but a low-thrust tug could, in a few months, gently push it to the cold, stable point a million miles from Earth. By the same means it could be brought back for repairs when necessary. The alternative is to leave the observatory in the cold, and to send out cryogenic robots to fix it *in situ*.

These are exciting times. While we sort out the challenges of space telescopes, a golden era of discovery should come from the much higher resolution and sensitivity and data throughput of big ground telescopes and interferometers. Then from space we should be ready to open wide the relatively unexplored infrared spectrum which holds such promise for discovery, from the furthest reaches of the Universe to planets of the nearest, ordinary stars.

Roger Angel is in the Department of Astronomy at Steward Observatory, University of Arizona, Tucson, Arizona 85721, USA.

1. Harwit, M. *Cosmic Discovery: The Search, Scope and Heritage of Astronomy* (Basic Books, New York, 1981).
2. Angel, J. R. P. & Fugate, R. Q. *Science* **288**, 455–456 (2000).
3. Tyson, J. A., Wittman, D. & Angel, J. R. P. in *Gravitational Lensing: Recent Progress and Future Goals* (eds Brainerd, T. & Kochanek, C.) ASP Conference Series (The Astronomical Society of the Pacific, in the press).
4. Angel, J. R. P., Cheng, A. Y. S. & Woolf, N. J. *Nature* **322**, 341–343 (1986).
5. Angel, J. R. P., Burge, J., Hege, K., Kenworthy, M. & Woolf, N. J. *Proc. SPIE* **4013**, 699–705 (2000).
6. Tyson, T., Kochanski, G. & Dell'Antonio, I. in *The Universe at Low and High Redshift* (eds Waller, W. H., Fanelli, M. N. & Danks, A. C.) 204–213 (American Institute of Physics, New York, 1997).
7. Burge, J. H., Cuerden, B. & Angel, J. R. P. *Proc. SPIE* **4013**, 640–648 (2000).

New physics with the Compact Linear Collider

John Ellis & Ian Wilson

Probing beyond the established picture of particle physics will require some radical re-thinking of accelerator designs. If accelerators are to reach the ever-higher energies that theorists would dearly like to see explored, the technological spin-offs of this engineering feat could be as surprising as the new subatomic physics.

Over the past decades, physicists' understanding of elementary particles and the forces that bind them has become enshrined in the picture known as the Standard Model. According to this model, the constituents of matter are particles of two basic varieties — quarks (which come in six varieties and, combined in threesomes, form protons and neutrons) and leptons (again, in six types, including the familiar electron, the muon, the tau and three types of neutrinos) (see Box 1). The forces between these particles are mediated by the exchanges of further particles, the 'messenger' bosons, the most familiar example of which is the photon, mediator of the electromagnetic force. The Standard Model is a highly successful theory, agreeing perfectly with all confirmed data from particle accelerator experiments, and describing accurately the characteristics of three of the four fundamental forces, the electromagnetic force and the strong and weak nuclear forces.

But the Standard Model has its limitations. As a theory, it is not entirely satisfactory, incorporating many arbitrary parameters. Moreover, it tells us nothing about gravity, the fourth and weakest of the fundamental forces; and there are hints from non-accelerator experiments observing neutrinos — ghostly particles that barely interact with other matter — that their behaviour cannot be fully accounted for in the Standard Model.

Therefore, in the twenty-first century the top priorities at particle accelerator laboratories around the world will be experiments probing beyond the Standard Model¹. Indeed, this is surely the only motivation for major new particle accelerators that is easy to explain to funding agencies, other scientists and the general public.

Problems beyond the Standard Model may conveniently be gathered into three main classes: those of mass, unification and flavour. Let us take these one at a time. What is the origin of the particle masses? One

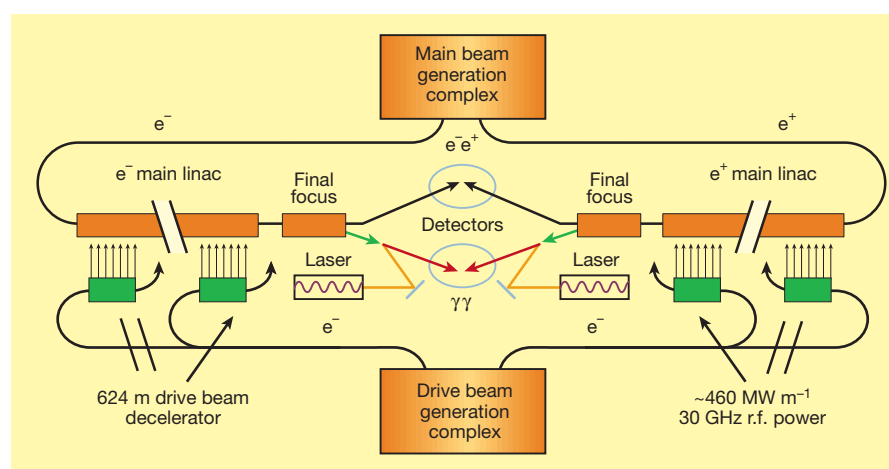


Figure 1 Overall layout of the CLIC complex.

possibility is that they arise through interaction with an all-pervading 'Higgs' field, which makes the vacuum behave like a superconductor. The question then is whether this field is composite, like the Cooper pairs of conventional superconductivity, or whether there is an elementary Higgs boson particle. If so, two related questions that arise are why the particle masses are so different (the mass of the top quark, for instance, is around 180 times that of a proton), and why they are so small. (One might well ask, 'Small relative to what?' The answer is that there is a natural scale of mass known as the Planck mass, obtained by combining the three universal constants — the speed of light c , Planck's constant, and G , the gravitational constant — and this mass is more than a billion billion times that of the proton.) Keeping the known particle masses small seems more natural mathematically if the Higgs boson is accompanied by a whole spectrum of new particles related to the known ones by supersymmetry: they would have the same interactions as the known matter particles, but be bosons rather than fermions.

Next is the problem of unification. The

electromagnetic and weak nuclear forces can be regarded as different manifestations of the same underlying phenomenon; their theories have been successfully combined into one 'electroweak' theory. Is there likewise a simple framework containing the strong, weak and electromagnetic interactions, and does it predict new observable phenomena such as proton decay and neutrino masses?

Finally, flavour. Why are there so many types of quarks and leptons? Looking at Box 1, we can see that there are six of each variety, not counting their antiparticles. How can one understand their weak mixing and the small, observed difference between matter and antimatter? Perhaps because the quarks are composite?

To probe ever deeper into the structure of matter, one needs to go to higher-energy experiments. Accelerator experiments essentially do two things: by smashing vastly accelerated particles into each other, they can either probe the particles' internal structure or create new particles. If one is looking at structure, higher energies provide improved resolution; if one wishes to create new particles, higher energies can create

heavier species (remember Einstein's famous equation $E = mc^2$). There are good reasons to expect a wealth of new physics in the teraelectronvolt range (1 TeV is 10^{12} eV, about 1,000 times the energy required to make a proton), in particular that connected with the origin of particle masses. This new physics might include a Higgs boson, but most physicists would expect the new physics to be more complex, perhaps including the new spectroscopy of supersymmetric particles mentioned above, or perhaps something even more exotic. (At the time of writing, there is speculation whether a Higgs boson may have been produced already at the Large Electron-Positron Collider (LEP) at the European Organization for Nuclear Research (CERN); if true, it suggests that supersymmetric particles might be 'just around the corner'².) Another possibility is that there might be extra dimensions of space-time that might show up in particles' internal structures, or be associated with new particles produced in the TeV energy range.

CERN is fortunate to have the Large Hadron Collider (LHC), a proton-proton collider with a centre-of-mass energy of 14 TeV, as an approved project. The LHC is currently under construction and scheduled for completion in 2006, which will provide a first glimpse of any new physics at energies up to about 1 TeV (in collisions between complex, multi-quark particles, not all of the energy is available for creating new particles). What it will find cannot be foreseen, but it is impossible to expect that experiments at the LHC will answer all the questions concerning this new physics. For example, particle physicists are likely to want more information about any kind of Higgs boson than the LHC can give us. Moreover, if nature has chosen supersymmetry, it can be expected that the LHC will reveal a number of supersymmetric particles but not all of them. And if the mechanism for generating particle masses turns out not to be an elementary Higgs boson but some new strong interaction, the hints that the LHC would provide should be followed up by other experiments.

Many of the open questions may be addressed best by a lepton-antilepton collider — one in which, for instance, electrons collide with their antiparticles the positrons, annihilating each other in a burst of pure energy, which in turn can create new particle-antiparticle pairs. In such a machine, all the centre-of-mass energy may be made available for the collisions between elementary particles, the experimental environment is relatively simple and clean, and all charged particles are produced democratically with similar cross-sections. Proton colliders and lepton colliders are complementary: the $W^\pm$ and Z^0 bosons were discovered in the Super Proton Synchrotron

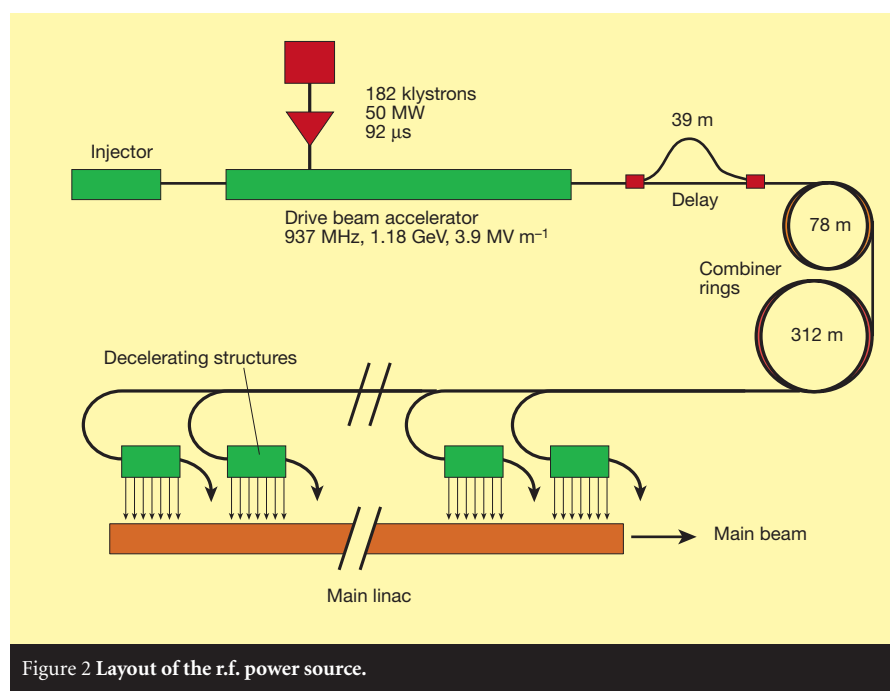


Figure 2 Layout of the r.f. power source.

(SPS) proton-antiproton collider, but it was LEP which made possible precision measurements of their properties and detailed tests of the standard electroweak theory.

Various laboratories are proposing electron-positron colliders with maximum centre-of-mass energies around 1 TeV, including SLAC (Stanford Linear Accelerator Center) and Fermilab (Fermi National Accelerator Laboratory) in the United States (1 TeV)³, DESY (Deutsches Elektronen-Synchrotron) in Germany (0.8 TeV)⁴ and KEK (High Energy Accelerator Research Organization) in Japan (1 TeV)⁵. Such machines would be able to explore in detail the properties of any relatively light Higgs boson and have a chance of producing lighter supersymmetric particles, but would probably not be able to explore all the supersymmetric spectrum, or study in detail any new strong interactions. Complete coverage of these issues, and hence full complementarity with the LHC, probably requires a lepton-antilepton collider with a centre-of-mass energy of 2 TeV or more.

This is the objective of the Compact Linear Collider, or CLIC as it is known. Plans for CLIC have been under way for several years at CERN, in collaboration with other accelerator laboratories⁶.

As we will describe later, the CLIC study team is proposing a new scheme of beam acceleration to enable electrons and positrons to be collided at energies ranging possibly from 0.2 TeV (for some overlap with LEP) up to a maximum of about 5 TeV (in stages). A road map has been drawn up to complete the research and development (R&D) necessary over the next several years to demonstrate the technical feasibility

of a 3-TeV centre-of-mass collider. The alternative of colliding muons and antimuons at comparable energies is also being explored at CERN and elsewhere⁷. In particular, CERN is conducting studies of the intense proton source that would be needed for such a muon collider. Bringing muons into collision at high energies would, however, require considerable R&D on cooling and other issues, so construction of any such high-energy muon collider is almost certainly on a longer timescale than CLIC.

The conceptual design of CLIC

One of the bugbears of high-energy electron-positron machines is synchrotron radiation: that is, the radiation emitted whenever beams of charged particles change direction. To combat the resulting energy loss, the circumferences of circular machines must increase rapidly with the energy. Scaling up CERN's present LEP accelerator, which is 27 km in circumference, one sees that a circumference of thousands of kilometres would be required to reach a centre-of-mass energy of 3 TeV. For this reason, future electron-positron colliders will be designed as linear machines.

CLIC's proposed overall layout is sketched in Fig. 1. The particles are accelerated to high energies by very high on-axis electric fields produced by the radio-frequency (r.f.) accelerating structures. Although the parameters correspond in terms of energy and luminosity to the desires of the physics community, there is a price to bear: to achieve the required luminosity requires the beam sizes at the interaction point to be very small (in the nanometre range), and the resulting very intense beam-beam interaction creates a spread in

Box 1 Glossary of terms: particles, forces and accelerators

Bosons. Particles of integer spin that carry forces.

CLIC decelerating structure. Device that generates microwave power when driven by an intense bunched electron beam.

Electromagnetic force. Holds atoms together by acting between charged particles, mediated by massless photons.

Fermions. Particles of half-integer spin that compose matter, comprising quarks and leptons.

Flavours. Different types of matter particles.

Gravitational force. Long-range, believed to be mediated by the massless graviton, whose existence has not yet been confirmed by experiment.

Hadrons. 'Heavy' particles that interact via the strong nuclear force, such as the proton and neutron, formed from combinations of quarks.

Higgs. A hypothetical field, believed to permeate the universe, providing masses for all the fundamental particles. Carried by the Higgs boson, whose existence has not yet been confirmed by experiment, although hints have recently been observed at LEP.

Klystron. High-power source of radio-frequency energy.

LEP (Large Electron-Positron Collider).

Accelerator colliding electrons and positrons at energies up to 105 GeV each. Housed at CERN in a tunnel with a circumference of about 27 km.

Leptons. 'Light' particles that do not interact via the strong nuclear force. They come in six types (flavours): the electron and its heavier charged cousins, the muon and tau, together with their three neutrinos.

LHC (Large Hadron Collider). Accelerator colliding protons at energies of 7 TeV each, to be installed in the tunnel previously holding LEP.

Linac. Linear accelerator in which large alternating voltages are used to accelerate charged particles in

a straight line; the concept dates back to the 1920s.

Quarks. The basic constituent particles of the proton, neutron, pion and stranger relatives. Interact via the strong nuclear force. They also come in six types (flavours): up and down, strange and charm, bottom and top, with increasing masses.

Radio-frequency accelerating structure.

Device that produces strong on-axis electric fields when powered by a high-power source of microwaves.

SPS (Super Proton Synchrotron). Accelerator used to collide protons and anti-protons at energies up to 540 GeV.

Standard Model. The theoretical description of the forces of nature and elementary particles. It describes the fundamental forces between matter particles via the exchanges of bosons. Tested to high accuracy by experiments at the LEP accelerator at CERN, and elsewhere.

Strong nuclear force. Holds protons and neutrons in the atomic nucleus, and quarks inside protons and neutrons, mediated by massless gluons.

Supersymmetry. A mathematical theory which suggests that, for every known type of particle, there should be a heavier 'supersymmetric partner', with identical internal properties, but different spin. Hypothetical as yet, the counterparts to the bosons are fermions with names ending in 'ino' (for example, chargino, photino, gluino) and those of the leptons and quarks are bosons whose names have an initial 's' (selectrons, squarks).

Weak nuclear force. Short-range, responsible for beta decay of radioactive nuclei, mediated by heavy W and Z bosons. Their existence was confirmed experimentally in 1983, and they have been studied in detail at LEP during the 1990s.

scheme involving the manipulation of intense electron beams to achieve both r.f. multiplication and power compression⁸. The scheme involves the following steps. First, generate the electron beam, using either a thermionic gun or a laser-illuminated photocathode. The second option is particularly demanding because it requires the generation of a huge charge from the photocathode (750 μC during 92 μs), and an average laser power of 860 W in the infrared to give 23 W in the ultraviolet. This is well beyond what can be achieved at present. Next, accelerate this beam to 1.2 GeV using a fully loaded linac operating at a relatively low frequency (937 MHz). The description 'fully-loaded' means that the beam takes almost all the energy (96%) from the structure, which is good for efficiency but makes the linac more difficult to operate. The r.f. power for the low-frequency linac is provided by about 200 50-MW klystrons. These klystrons would normally be very classical, were it not for the fact that the pulse length required is about 100 μs . The development of such long-pulse klystrons is one of the many challenges of the CLIC scheme, and is being studied by European industry.

The electrons in the beam at this stage are in bunches spaced 64 cm apart, which is a requirement to be able to accelerate the beam with the available low-frequency technology. The beam has all the required bunches for the 22 drive beams needed for one linac of a 3-TeV collider. But the spacing has to be reduced to 2 cm, to be able to generate the 30 GHz power, and this is done by funnelling the beam in compressor/combiner rings. The trick is to take successive parts of the beam and interleave them to reduce the bunch spacing (this is called frequency multiplication), and it also results in an increased line density of the charge, which is effectively pulse compression. There are three stages of frequency multiplication in the CLIC scheme (the delay line and two combiner rings) giving a total compression factor of 32.

The layout of the r.f. power source is shown in Fig. 2. The resulting long train of separated drive beams is then sent into the main tunnel. Each drive beam is used to provide r.f. power for a 625-m section of the main linac, after which the beam is sent to a dump and the following drive beam takes over, as in a relay race.

If this scheme is considered as a black box, 1 GHz r.f. power is put into one side, and pulse-compressed 30 GHz r.f. power comes out from the other side. An attractive feature of this scheme is that generating more or fewer drive beams to power a higher- or lower-energy collider requires only a longer or shorter modulator pulse, but the number of klystrons does not change. This so-called 'two-beam' acceleration scheme is generally

beam energy owing to photon radiation.

The CLIC scheme⁶ has two key features that distinguish it from other, lower-energy linear collider studies⁸. The first is the operating frequency of the accelerating structures in the main linacs. To limit the length and cost of these linacs, high accelerating fields are mandatory, and experience has shown that these can be obtained (with conventional acceleration mechanisms) only by operating at a high frequency. CLIC has therefore chosen to operate with a radio frequency of 30 GHz, in the hope that it can achieve accelerating gradients as high as 150 MV m^{-1} . The resulting total length of 37.5 km for a 3-TeV collider is comparable to the circumference of CERN's present LEP accelerator.

The second distinctive feature of the CLIC scheme is the way in which it generates the r.f. power that produces the accelerating

field. High-intensity but low-energy 'drive beams' of electrons run parallel to the main beam, and the power is extracted from these by specially designed decelerating structures. This is a particularly attractive feature, because the energy for r.f. power production is in the electron beam, which can be transported over long distances with very small losses, and the r.f. power is generated locally only where it is required. In fact, the transport distance from the drive linac to the main linac is then only about 60 cm. It should be noted that to generate an accelerating gradient of 150 MV m^{-1} requires the production of peak pulsed powers of 460 MW per metre length of the linacs but only for a very short time of 120 ns.

Generating the intense CLIC electron drive beams is far from straightforward. Years of study at CERN have resulted in a

acknowledged to be the only cost-effective way of building multi-TeV electron–positron colliders. A virtue of this scheme is that it leads to a simple tunnel layout. The tunnel has only to house the two linacs, supported on a solid concrete base the width of a table-top (albeit a long one), and the various beam transfer lines together with their beam-focusing quadrupole magnets. A tunnel diameter of 3.8 m, the same as LEP, would be adequate.

Turning our attention now to the interaction point, we see that a steep technical challenge is imposed. The interaction rates for the processes of interest generally decrease as the square of the centre-of-mass energy. So the collision rate (luminosity) of a high-energy collider should be much greater than that of LEP, which reached around 10^{32} collisions $\text{cm}^{-2} \text{s}^{-1}$ at a centre-of-mass energy near 200 GeV. Hence the request to the accelerator designers is for a machine capable of around 10^{34} collisions $\text{cm}^{-2} \text{s}^{-1}$ at a centre-of-mass energy of 1 TeV (or 10^{35} collisions $\text{cm}^{-2} \text{s}^{-1}$ at 3 TeV). Therefore, the CLIC study team has been forced to select beam parameters that are at the edge of what is considered to be technically feasible. The vertical beam size, for example, has to be focused down to 1 nm at the interaction point. This implies the generation of very small emittance beams using damping rings with a performance more than five times better than has been achieved in the present state-of-the-art damping ring at KEK in Japan⁹. (Emittance is a measure of the density of the particles in the transverse plane of the beam.) This emittance has then to be preserved during the acceleration process, in spite of perturbing transverse deflecting fields, which increase as the cube of the frequency. Potentially this is a large disadvantage of using 30 GHz, but it can be surmounted if other parameters are chosen carefully. Indeed, it has been shown that the same beam stability can be achieved in a high-frequency linac as in a low-frequency linac, by making a judicious choice of parameters according to general scaling laws¹⁰.

This analysis of beam stability assumes that the accelerating structures can be designed to suppress any generated perturbing fields by a factor of 100 in a time of 0.7 ns. Such a structure has been developed at CERN¹¹, and tested with an electron beam at SLAC¹². It consists of a regular copper travelling-wave accelerating structure with each of its 150 cells damped by its own set of four radial waveguides.

Working at 30 GHz provides huge technological challenges. The fabrication of the 30-GHz accelerating structure, for example, requires the machining of copper component parts to a precision of 1 μm , using state-of-the-art lathes developed initially for the mass-production of contact

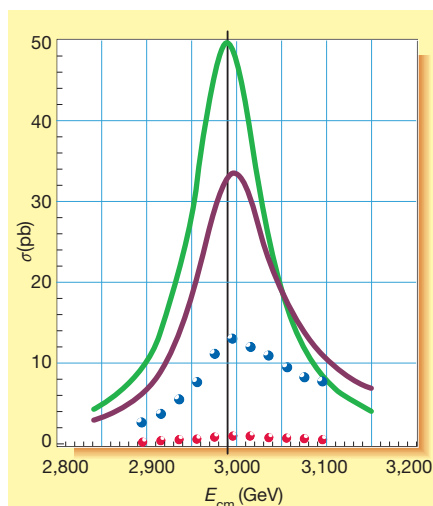


Figure 3 Simulation — preliminary as yet — of Z' production at CLIC. The vertical axis measures the event rate (in picobarns), and energy is plotted on the horizontal axis. Solid lines indicate the rates expected without photon radiation (green) and with the minimum amount of radiation possible (purple) in an idealized collider. The simulated data points are for quark–antiquark production (upper blue trace) and lepton–antilepton production (lower red trace), including the photon radiation expected at CLIC. Both processes are clearly visible.

lenses. Another consequence of choosing 30 GHz is that the beam aperture, which is inversely proportional to frequency, is only about 3.5 mm in diameter along the 13.75 km length of the linac. Seen from the outside this is stunningly small, but should provide adequate space when viewed by the electron beam.

To keep the perturbing fields within acceptable limits imposes another stringent condition. The components in the linacs have to be aligned transversely and maintained in position to within 5–10 μm along typical lengths of about 200 m. This requirement means that a static one-time alignment of the linacs is insufficient, and imposes an active-alignment system that automatically reacts to the continuous drifts of the components. Such a system has been developed at CERN using state-of-the-art technology including 0.1- μm resolution stepping motor drives, capacitive position monitors (in which the position of a conducting wire between two capacitive plates changes) and hydrostatic levelling systems, and has been used in a real accelerator environment to maintain components in position to within 1–2 μm .

Colliding and maintaining 1-nm beams in position is clearly far from easy. It imposes a jitter tolerance of 0.2 nm on the final focusing elements of the beam delivery

system. Stabilization of magnetic quadrupoles to this level is a huge challenge and demonstrating its technical feasibility is a priority. Two stabilization schemes are being considered: the first is an optical anchor to the nearby bedrock using laser interferometers, the second is based on inertial sensing and compensation using piezo-movers.

Tests and prospects

Preparing such a novel accelerator concept requires years of R&D before a concrete project can be proposed. The ideas proposed by the CLIC study team are currently being probed in a series of test facilities, and outlines prepared for future studies.

The first test facility, operating from 1990 to 1995, demonstrated the feasibility of two-beam power generation, albeit at lower field gradients than will ultimately be required. A second test facility is now being operated¹³. It consists essentially of a 6-m long, two-beam test accelerator driven by an intense 40-MeV and 3-GHz bunched beam. The 30-GHz part of this facility is equipped with an active-alignment system capable of alignment to a precision of a few micrometres. This facility is being used as a test bed to study the generation of short intense electron bunches, to develop diagnostic equipment, and to generate high-power 30-GHz r.f. pulses for high-gradient testing of accelerator components.

A new facility (CTF3) is to be constructed in collaboration with LAL (Laboratoire de l'Accélérateur Linéaire, France), LNF (Laboratori Nazionali di Frascati, Italy) and SLAC, which would test all major parts of the CLIC r.f. power scheme¹⁴. To reduce cost, it is based on the use of 3-GHz klystrons and modulators recuperated from the LEP Injector Linac (LIL). Construction will start after the closing of LEP, and its study programme should continue until about 2005. At that time, a conceptual design report could be made for CLIC1, a prototype CLIC test accelerator consisting of one complete drive-beam and acceleration unit. This test accelerator would be capable of accelerating a beam to 75 GeV and would, if successful, provide a convincing demonstration of the technical feasibility of the CLIC two-beam scheme.

What might CLIC see?

For lower-energy electron–positron colliders, detailed studies of the physics abound. These give us clues to the physics that CLIC might have to offer. A more comprehensive study of CLIC's specific possibilities has recently been initiated at CERN.

The preliminary studies indicate clear complementarity between the physics prospects for the LHC and CLIC. In particular, CLIC has unique prospects for finding any new particles that do not have strong interactions. One issue is that, although

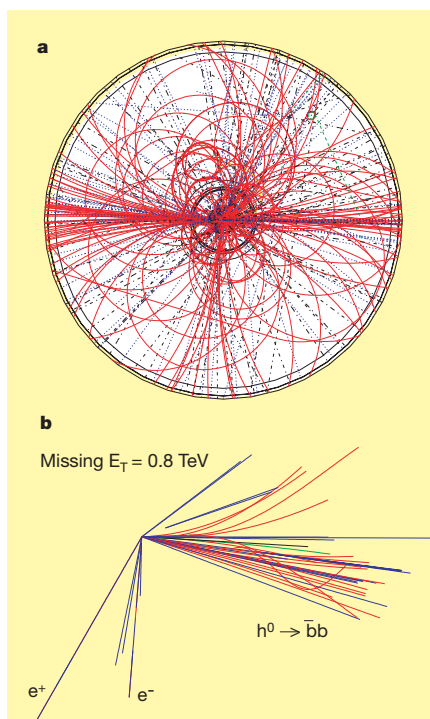


Figure 4 New physics that CLIC might explore. **a**, Simulations of charged-Higgs-boson pair production and decay. Charged Higgs particles are unstable and decay into jets of other charged particles that leave tracks in an experimental detector. The jets need to be combined to identify the charged Higgs bosons. **b**, Simulations of electron pair production and decay. One of the produced selectrons decays directly into an electron and an unseen neutral supersymmetric particle, whereas the other decays via a cascade including a neutral Higgs boson.

CLIC is much less wasteful than proton–proton colliders, compared with lower-energy electron–positron colliders it loses a relatively large fraction of the beam energy in the form of radiation from the initial-state particles. It is important to be satisfied that the resulting beam-energy spread permits experimental investigation of the full richness of the available physics¹⁵.

One of the first examples of new physics to be considered is a new Z' particle akin to the Z boson, the known carrier of the weak interaction, which arises generically in models with new strong interactions or large extra dimensions. It might show up as a ‘resonance’ — that is, a bump in the plot of the interaction rate against beam energy when the latter is tuned to the Z' mass. Fig. 3 shows a simulation of what such a Z' particle might look like at CLIC. Photon radiation reduces the total cross-section by a larger factor than for the Z at LEP, but the resonance is clearly visible. CLIC could easily be a Z' factory, just as LEP was a Z factory in its initial stages.

In more recent years, LEP has studied the

production of pairs of W particles, and has also been looking for a neutral Higgs boson, produced in association with the Z boson. We do not yet know what new particle thresholds CLIC might be called upon to explore, but two examples are shown in Fig. 4. Figure 4a shows a simulation of an event in which a pair of heavy charged Higgs bosons, predicted in supersymmetric models, have been produced, and Fig. 4b shows a simulation of an event in which a pair of selectrons have been produced (these are the proposed supersymmetric counterparts to electrons), one of which decays into a lighter neutral Higgs boson. These events give some flavour of the richness of the new physics that would be accessible to CLIC, including many topics that extend beyond the reach of the LHC.

As already mentioned, the interaction rates for many interesting processes such as these decrease as the square of the centre-of-mass energy — hence the request to the designers for a machine capable of 10^{34} collisions $\text{cm}^{-2} \text{s}^{-1}$ at 1 TeV, and higher still at 3 TeV.

New technologies

What about the technological spin-offs from CLIC? We touched earlier on several of the leading-edge technologies required. Many need considerable R&D, in collaboration with other accelerator laboratories and European industry.

Although the CLIC design has been optimized for a centre-of-mass energy of 3 TeV, the collider could start operation at a lower energy and then be upgraded in stages. These upgrades could be made without major modifications. The basic CLIC scheme could of course be adapted to generate r.f. power at lower frequencies and indeed SLAC is studying a two-beam scheme at 11.4 GHz. The cost advantage of using the two-beam scheme becomes more pronounced as the energy increases and there is a consensus within the community that future multi-TeV colliders will probably be based on the two-beam scheme. In the range 0.5–1 TeV, it is difficult to say how the cost would compare with the lower-frequency classical acceleration schemes currently being proposed in the United States³, Germany⁴ and Japan⁵.

More generally, one may wonder whether the principle of a high-gradient (long) table-top electron linac might be of interest in other applications. But we should not forget that there is a price penalty to be paid compared with conventional systems at very low energies, because to produce beams of only a few GeVs requires the building of the entire drive-beam generation complex.

Existing electron accelerator technologies are already used extensively for synchrotron radiation sources, and some of the linac designs now being proposed offer exciting prospects for free-electron lasers and X-ray sources. This is an obvious use of the beam

generated by the CLIC1 facility, and studies in this direction are starting.

As mentioned above, for CLIC to be successful, its subsystems require technological developments that may find applications elsewhere. Examples include the work going into high-power diode-pumped lasers for CLIC, which could benefit the development of laser-produced plasma sources of extreme ultraviolet light that are used in industry for photolithography. The development of the klystrons for the drive beam is based on multi-beam technology that is relatively new for European industry, and will bring benefits of improved efficiency and lower voltage which will almost certainly find other applications. Furthermore, if the technological challenges of high-precision fabrication, alignment and stabilization can be met, all of these could lead to significant technological spin-off.

Other possible applications of the CLIC technology are under study, and may be found in the least obvious places. When particle accelerators were first developed in the 1940s, nobody foresaw their medical applications (for example, positron emission tomography scanning or hadrotherapy), or their applications to other areas of science (synchrotron radiation was initially regarded as an annoyance). Still less did anybody foresee that accelerator experiments would drive the development at CERN of the World-Wide Web. A recent unexpected application of accelerator technology concerns the flat ribbon-like non-evaporable getter pumps developed for the LEP vacuum system, which will almost certainly be used for flat-panel displays, electron tubes and vacuum-insulated devices. Because past breakthroughs in accelerator technology have found unforeseen uses in other fields besides particle physics, we expect that CLIC will surely do the same.

John Ellis is in the Theoretical Studies Division, and Ian Wilson is the Deputy CLIC Study Leader, CERN, CH-1211 Geneva 23, Switzerland.

1. Ellis, J., Keil, E. & Rolandi, G. Report CERN-EP/98-03 (1998).
2. Ellis, J., Ganis, G., Nanopoulos, D.V. & Olive, K. A. Los Alamos e-print archive, hep-ph/0009355.
3. The NLC Design Group. LBL-PUB-5424, SLAC Report 474, UCRL-ID-124161 <http://heplibw3.slac.stanford.edu/accel/nlc/zdr/> (1996).
4. Brinkmann, R., Materlik, G., Rossbach, J. & Wagner, A. (eds) Report DESY-97-048 <http://www-library.desy.de/preprints.html#reports> (1997).
5. The JLC Group. KEK Report 92-16 <http://www-lib.kek.jp/lists/pubadd.html> (1992).
6. CLIC Study Team. Report CERN/PS 2000-047 (2000).
7. Autin, B., Blondel, A. & Ellis, J. (eds) CERN Report 99-02 (1999).
8. Braun, H. et al. CERN Report 99-06 (1999).
9. Urakawa, J. Contribution to the 1997 Particle Accelerator Conference (PAC'97), 12–16 May 1997, Vancouver <http://www.triumf.ca/pac97.html>.
10. Delahaye, J. P. et al. Report CERN-PS/97-51 (1997).
11. Wilson, I. & Wuenisch, W. Report CERN/PS 2000-057 (2000).
12. Wilson, I. et al. Report CERN/PS 2000-044 (2000).
13. Bossart, R. et al. Report CERN-PS/98-60 (1998).
14. CLIC Study Team. Report CERN/PS 99-47 LP (1999).
15. CLIC Physics Study Group. <http://cliphysics.web.cern.ch/CLICphysics/>.

Lab automation ideas

Automated metabolism studies, and software to go under the microscope.

WinHg

Leeman Labs www.leemanlabs.com

Software for mercury analysers

This new software package can be used with Windows 95, 98 and NT platforms to operate the PS200, PS200II and Hydra series of mercury analysers. It features a graphical user interface and enhanced file maintenance, with improved report generation allowing creation of up to 3,000 records. An automated quality control sequence validates analyser performance before running samples and can be individually customized. Audiovisual aids for routine and scheduled maintenance tracking are provided, while enhanced protocol features allow calibration of the analyser with up to 10 standards. Customer support is provided via remote communications software.

Reader Service No. 100

Discovery SE

Kendro Laboratory Products

www.kendro.com

Silent spin

Offering top speeds of 100,000 r.p.m., this ultracentrifuge is claimed to be the smallest floor model on the market, and to be whisper-quiet in operation. Routine spins are set up and monitored using a keypad and dual LCD display showing both set and run conditions. A screen zoom function allows separations to be viewed from across the laboratory. Other features include automatic r.c.f./r.p.m. mode for setting maximum speed or desired g force, real-time control for delayed start/stop runs, an on-line rotor catalogue detailing key specifications for Sorvall, Beckman and Kontron rotors, a 20-program memory, and a run scheduling function for centrifuge reservations.

Reader Service No. 101

MBS 1 filtration system

Schleicher & Schuell www.s-and-s.com

Microbiological filtration system

Designed for use in quality control procedures, the three-part MBS 1 filtration system consists of funnel, apparatus, and membrane. The funnel dispenser and membrane butler are combined so that when the funnel is taken out of the dispenser, the butler automatically makes a membrane available that is removed using forceps from the sterile pack, and is then ready for use. This system is designed to make for easier validation and to achieve a time saving of approximately 30 seconds per test. Other features include a



Test case: S&S microbiological filters.

large funnel capacity for foaming liquids and a safe-sealing mechanism.

Reader Service No. 102

SPE Dry Dual

Jones Chromatography

www.joneschrom.com

A lot of hot air

This 96- and 384-well plate sample concentrator features side-by-side needle assemblies for simultaneous drying of samples in two microplates. Heated gas, typically nitrogen, is delivered from needle assemblies located above and below the microplate,



Concentrate: SPE Dual Dry, for 96 or 384 wells.

enhancing evaporation of aqueous solvents and resulting in rapid sample concentration. The assemblies are removable for cleaning, and the plate holder can be adjusted to any height from the upper assembly.

Reader Service No. 103

Chromolith

Merck

www.merck.de

HPLC columns specialized for high throughput

Chromolith Performance and Chromolith SpeedROD high-performance liquid chromatography columns are described as light, slim, and up to 10 cm long, with a porous silica gel skeleton and a coating of black plastic. They are said to be capable of separating a mixture containing 33 different pesticides in under 15 minutes, compared to one hour with conventional HPLC columns. More complex mixtures can be separated by joining Chromolith Performance columns together, up to a length of one metre.

Reader Service No. 104

LabManager iLIMS

Beckman Coulter www.beckmancoulter.com

Lab information via the web

This laboratory information management software package allows users access via a web browser or client/server using a Microsoft Outlook-styled interface. The system is a native Oracle8 application providing complete sample, test and data management on a Windows NT/2000 or HP9000 server. The software supports both Microsoft Visual Basic to automate processes and customise applications, data formats and calculations, and Seagate's Crystal Reports tool for custom report design.

Reader Service No. 105

Data management

Thermo LabSystems

www.thermolabsystems.com

New developments in microarray data handling and visualization

A novel A2D (analogue to digital) chromatography data system for single workstations and the latest interface developments between Thermo's laboratory information management systems and SAP's R/3 Enterprise Resource Planning System and the OSI PI Plant Information system are among the developments that the company will be demonstrating at the Pittcon exhibition in New Orleans on 4-8 March 2001. Presentations on study management software

new on the market

and application service provision will be given. Information on secure software delivery over the Internet and electronic laboratory notebooks will also be on offer.

Reader Service No. 106

Axiocision Mark & Find

Zeiss www.zeiss.de

Mouse map

Marking and relocation at the click of a mouse is promised by the makers of this software module. It can be used to mark positions on microscope slides and to reposition marked objects for further processing, and is suitable for motorized and encoded stages. A database is available for case-related management of data from several specimens and position lists. The recorded results can be allocated to different classes. Summary information is provided by a display of the position co-ordinates in an overview of the specimen (coloured marking of the positions) or as a list.

Reader Service No. 107

Multidose G3

Zymark www.zymark.com

Automated workstation

Zymark's Multidose G3 provides spectrophotometric data acquisition, results calculation, database storage and a comprehensive audit trail. It is fully compliant with 21 CFR Part 11, can be integrated with a wide variety of third-party device drivers and can be used to create customized reports to accommodate individual laboratory templates or forms. Access is controlled by a system administrator that assigns various privilege levels to multiple users. As with other MultiDose products, the G3 can run multiple methods in a single, unattended sequence, making it particularly suitable for use in dissolution methods development labs, stability testing groups, or any busy lab running many different products.

Reader Service No. 108

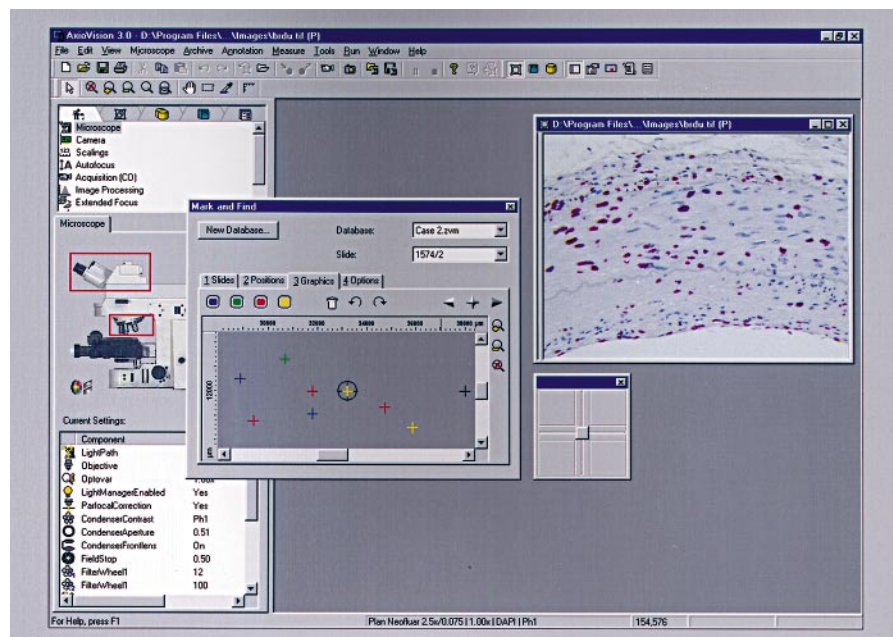
CellQuest Pro 4.0

Becton Dickinson www.bdfacs.com

Snappy software

This software release is an expanded and updated version of the company's CellQuest 3.3. Features include user-defined calculations for quick and easy statistics; improved tools for managing acquisition, including random access of samples; export of graphics and entire experiment documents in a variety of formats; a PDF facility for creating custom reports, and a new snap-to-region tool for faster and more accurate data analysis. A demonstration can be seen at www.bdfacs.com. The software is available either as an upgrade to CellQuest 3.3 or as a full package.

Reader Service No. 109



Slide show: Mark&Find software automatically repositions objects on specimens under the microscope.

MetaboLynx 3.5

Micromass www.micromass.co.uk

Metabolite screening and identification

While automated LC-MS/LC-MS-MS provide the classic tools with which to explore both metabolic fate and the kinetics of biotransformations, data interpretation has remained a rate-limiting step. The MetaboLynx 3.5 application manager, however, detects peaks in an LC-MS data file resulting from in vitro or in vivo biotransformation and provides rapid metabolite identification. The system detects putative biotransformations without the need for operator-supplied predictions of metabolic fate. It automatically runs samples sched-

uled for LC-MS analysis and processes the resulting data using a data browser that allows chromatographic and mass spectroscopic evidence to be rapidly reviewed locally or remotely via a secure corporate network. Major features include interrogation of LC-MS data for targeted metabolites; correlation of metabolized sample data with a control sample; pinpointing of untargeted biotransformations; generation and execution of method files; integration of chromatographic peaks for semi-quantification and electronic reporting/distribution of results.

Reader Service No. 110

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

BACs • YACs • PACs cDNAs

Research Genetics provides clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

Research Genetics

U.S. or Canada 800-533-4363

U.K. 0-800-89-1393

FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

Feeling short changed by your Sequences?

COMPLEMENT GENOMICS offers.....

- ☐ More bases for your money
- ☐ Quality values for each
- ☐ Quick turnaround

Contact us **today** to discuss your requirements

www.complementgenomics.co.uk

Tel:+44 (0) 191 516 6500

Fax:+44 (0) 191 516 6501

READER ENQUIRY NO. 40

READER ENQUIRY NO. 54